2014

A local structure graph model for network analysis

Emily Taylor Casleton
Iowa State University

A local structure graph model for network analysis

by

Emily Taylor Casleton

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Mark S. Kaiser, Co-Major Professor

Daniel J. Nordman, Co-Major Professor

Petruţa C. Caragea

Arka P. Ghosh

Max D. Morris

Alyson G. Wilson

Iowa State University

Ames, Iowa

2014

DEDICATION

I would like to dedicate this thesis to my husband, Dave, without whose support, patience, encouragement, and whiteboard discussions I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, my advisers, Dr. Dan Nordman and Dr. Mark Kaiser, for their guidance, patience, and support. I have learned a lot from working with them, and I feel I am a much better statistician now than when we began this process with that first trip to Albuquerque. I would like to thank my committee members for their time and patience: Dr. Max Morris, Dr. Petruţa Caragea, and Dr. Arka Ghosh. Also, I would like to thank Cindy Phillips and the team at Sandia National Laboratory for motivating the research in network analysis and for funding this work through a Laboratory Directed Research and Development grant.

I would like to thank others who have helped me succeed at Iowa State. To my cohort of fellow graduate students, I would not have made it through the first few years without you. I have heard that the people you attend graduate school with become your colleagues, and I hope I have the opportunity to work with each of you in the future. And to my pod-mates, Adam and Dan, your advice, elaborate motivation, and company on coffee runs has helped me tremendously. Most importantly, I would like to thank the Department of Statistics for encouraging a collegial environment and their continued support and development of graduate students.

Last, but not least, I would like to thank Dr. Alyson Wilson who played a large part in my decision to attend ISU and is the reason I got involved in the project that lead to this thesis. I also greatly appreciate her mentoring and advice throughout my graduate career.

ABSTRACT

The statistical analysis of networks is a popular research topic with ever widening applications. In this work, we introduce a new class of models for network analysis, called local structure graph models (LSGMs). The approach specifies a network model through local features and allows for an interpretable and controllable local dependence structure. In particular, LSGMs are formulated by a set of full conditional distributions for each network edge, e.g., the probability of edge presence/absence, which depend functionally on neighborhoods or subcollections of other network edges. Hence, LSGMs correspond to a type of Markov Random Field (MRF) model applied to graph edges. The modeling features and interpretation of LSGMs are demonstrated through several numerical studies and illustrated through a network data example involving tornado occurrences. LSGMs are also shown to provide an alternate specification of another popular class of models for random graphs, belonging to exponential random graph models (ERGMs), which specify a model through a joint distribution on the entire collection of graph edges. An ERGM induces conditional distributions and neighborhoods, rather than explicitly defining them as in the LSGM approach. As one consequence of its conditional specification, LSGMs have the advantage of allowing direct control and separate interpretation of parameters influencing large-scale (e.g., marginal means) and small-scale (i.e., dependence) structures in a graph model. This is possible with LSGMs through so-called *centered parameterizations* of MRF models, which ERGMs are shown to lack. The centered parameterization and conditional specification of LSGMs further provide important advantages in graph modeling when incorporating covariate information from nodes, as illustrated with two further network data examples. However, the centered parameterization was developed for MRFs under an assumption of pairwise-only dependence, meaning that dependence is modeled between pairs of dependent edges only. This particular dependence structure may be inappropriate for modeling network data that exhibit transitivity or a prevalence of triangles within the network, which

has been identified as an important feature of various networks. Consequently, the centered parameterization for MRFs is extended to account for triples of dependent edges in LSGMs. This extension then allows for the explicit modeling of transitivity in LSGMs, while retaining the same interpretable separation and control of large- and small-scale effects in a graph model and facilitating the use of covariate information. At the same time, the ability to model transitivity does not imply that this model feature should be commonly used or applied without cautious model diagnostics, which are currently lacking for graph models and for ERGMs in particular. By developing simulation-based model assessments for random graphs, we provide in-depth examinations and analyses of two commonly-used example networks, demonstrating that real network data may not, in fact, support the inclusion of transitivity in a graph model.

CHAPTER 1. INTRODUCTION

1.1 Background

Since the mid-1990's, there has been a research explosion in the area of network science. There are conferences, (e.g., the International Network for Social Network Analysis annual conference, the Intra-Organizational Networks conference, and the annual international conference on Advances in Social Network Analysis and Mining), network analysis centers (e.g., Duke Network Analysis Center, LINKS Center for Social Network Analysis at the University of Kentucky, and the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University), special issues of journals (e.g., volume 21, issue 4 of *Journal of Computational and Graphical Statistics* (2012), volume 21, issue 3 of *Journal of Statistical Software* (2008), and the inaugural issue of the journal *Network Science* appeared in 2013), and computational packages (e.g., Stanford Network Analysis Project, the `statnet` suite (Goodreau et al., 2008) and `ergm` (Hunter et al., 2008b) packages in `R`, SIENA (Snijders et al., 2006), and Statistical modeling of sOCial NETwork (StOCNET)).

However, network analysis predates this recent increase in interest. Visualization of graphs as points and lines, known as a sociogram, was introduced by Moreno in 1934 (Fienberg, 2012). One of the first social networks arose from an experiment conducted by Stanley Milgram in the early 1960s which measured the number of connections between two randomly chosen individuals through chain letters. Although a large majority of chains were never completed, Milgram discovered the median length of completed chains was only 6, inspiring the play and movie *Six Degrees of Separation* (Goldenberg et al., 2010). Probability models for networks can be traced to the Erdös-Rényi-Gilbert graph in 1959 (Kolaczyk, 2009) and the International Network for Social Network Analysis, an association for social-network researchers, was founded

over three decades ago in 1977 (INSNA, 2013). Although the value of networks has long been known, it is only recently that their potential has been realized due to the combination of recent advances in computational ability, which makes analysis of large networks possible, and the emergence of large, freely-available, and interesting networks, such as the Internet, Facebook, or the Wikipedia.

A network is defined by a set of nodes and the relations between them. Networks model relational data, or data with features that cannot be described by only examining the individual nodes (Handcock, 2003a). Complex patterns of connections and dependencies can be represented by a network and between a potentially large number of nodes. Due to this ability, applications from a wide variety of disciplines have been appropriately modeled as a network. For example, biologists have modeled the structural connectivity of brains with networks (Sporns et al., 2004; Simpson et al., 2011), zoologists use graph models to represent the social network of animals, such as dolphins (Lusseau, 2003), international relations between countries have been analyzed with networks (Hoff and Ward, 2005; Barigozzi et al., 2010), and epidemiologists use networks to understand the spread of a disease within a population (Groendyke et al., 2012). In addition, networks have represented the effect of disasters on fiber-optic networks (Neumayer and Modiano, 2010), the communications between a cell of a terrorist networks (Schweinberger and Handcock, 2012), the reliability of sampled data on a protein-protein interaction networks (Raftery et al., 2012), and the inter-organizational collaborations between rescue and relief organizations in response to the September 11, 2001 attack (Schweinberger et al., 2012).

This wide variety of applications from various disciplines demonstrates the flexibility of networks as a modeling tool. However, this diverse array of problems has led to a diverse array of solutions from a diverse array of disciplines. As one review aptly described, "Network science has no home" (Vivar and Banks, 2012). Many of the models or methods developed for the analysis of networks have been for a specific application or observed network, and thus most are very ad hoc and not appropriate for other types of observed networks (Vivar and Banks, 2012). This dissertation will present a newly-constructed class of models for the statistical analysis of observed graphs, local structure graph models (LSGMs), which were not developed

for a specific application, but rather is an application of the binary Markov Random Field (MRF) model to graph edges.

1.2 Overview

1.2.1 Literature Review

Network science is vast with published works appearing in a diverse array of fields. Due to extensive and broad range of existing literature, a comprehensive review is not possible, so the literature review presented in Chapter 2 will focus on the contributions from computer scientist/statistical physics and sociologists/statisticians; two subsets of network science that have contributed a considerable amount. The reasons for choosing these two areas is that motivation for this dissertation was a collaboration with a team of computer scientists at Sanida National Laboratory and it was desired to understand the current work relevant to our collaborators. A majority of statisticians working on network analysis have focused on social networks, and thus the goals of the literature in the area of sociology are similar to those of this dissertation.

1.2.2 The Local Structure Graph Model (LSGM)

The new class of models for network analysis, LSGMs, will be introduced in Chapter 3. There are two distinguishing features of a LSGM. The first is the specification, for each potential edge, of a conditional distribution, i.e., the distribution for the presence/absence of that edge given the outcomes of all other potential edges in the network. An explicit definition of dependent sets of edges, called neighborhoods, is the second characteristic of a LSGM. Markov dependence is assumed so that edges are conditionally dependent only on edges that belong to the same neighborhood. Two additional features of LSGMs, the ability to simply incorporate potential spatial information about nodes, and the definition of a "saturated graph," are also introduced. These features can help keep the potential sizes of LSGM neighborhoods manageable which helps avoid a common issue of model degeneracy. This chapter will also show that

a LSGM can be interpreted as an alternate method of specifying another model for random graphs, an exponential random graph model (ERGM).

1.2.3 LSGM with Higher-Order Dependence

A LSGM is based on an application of the binary MRF model to the edges of a network. A MRF model is generally used to analyze geo-referenced data because of its ability to incorporate spatial dependence. This dependence is often modeled between pairs of random variables only, an assumption known as pairwise-only dependence, which is sufficient to spatial applications. However, this assumption creates a limitation in specifying conditional distributions for graph edges in LSGMs where it may be necessary to incorporate dependence between triples of dependent edges. Another consideration is the parameterization of conditional distributions which can have an important effect on model parameter interpretation. LSGMs have the advantage of allowing direct control and separate interpretation of parameters influencing large-scale (e.g., marginal means) and small-scale (i.e., dependence) structures in a graph model due to the use of a so-called *centered parameterization*. However, this parameterization was also developed under the pairwise-only dependence assumption. Thus, Chapter 4 extends the centered parameterization to allow for higher order dependence and shows that the parameter interpretation advantage is maintained.

1.2.4 Importance of Transitivity

The extension introduced in Chapter 4 allows for a LSGM to explicitly model dependence between triples of dependent edges. For an incidence definition of dependence, the common dependence structure in network analysis where two edges that do not share a node are conditionally independent, the two topological features that lead to triples of dependent edges are 3-stars and triangles. Modeling transitivity, the prevalence of triangles within the network, has received a lot of attention due to the perceived prevalence in various networks of interest and intuitive scientific interpretation. For example, in a social network, transitivity is demonstrated through two individuals who are more likely to be friends when they share a common friend. Although the idea of transitivity is intuitive, effects that are modeled need to be supported by

the data structures. In Chapter 5, two networks which are commonly used as examples in the literature are studied in detail. These two particular networks were chosen because they allow for a comparison between too little and too much realized structure. Specifically of interest is if the data supports the inclusion of transitivity, or dependent triples of edges in each situation. Simulation-based model assessments, indicate that real network data may not, in fact, support the inclusion of transitivity in a graph model, indicating a need for cautious model diagnostics.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

The literature review presented in this chapter will examine a subset of the research published in the fields of computer science and statistical physics and sociology and statistics. These two areas have made considerable contributions to the advancement of network science. The chapter will not include an exhaustive list of the random graph models from these areas, as network science is vast and quickly evolving in all disciplines. Although there exists multiple literature reviews of network analysis (e.g., Vivar and Banks, 2012; Fienberg, 2012; Goldenberg et al., 2010; Salter-Townshend et al., 2012; Chakrabarti and Faloutsos, 2006), this review is unique in its classification of the network analysis techniques into algorithmic construction and probabilistic modeling.

Most modern models for random graphs arose from the often misinterpreted Erdös-Rényi graph model. Two specifications of this model were proposed at approximately the same time in 1959. In a series of papers, Erdös and Rényi (Kolaczyk, 2009) specified a random graph model where the number of nodes, n, and the number of edges, m, are fixed, and a uniform distribution is placed on all N possible graphs, where

$$N = \binom{\binom{n}{2}}{m}.$$

In the same year, Gilbert (1959) proposed his specification of the same model where the number of n nodes are fixed and edge formation occurs according to a constant, independent probability p for each of the $\binom{n}{2}$ pair of nodes. From this specification, the likelihood of a particular graph can be determined and is the binomial distribution. Unjustly, Gilbert's specification is often referred to as the Erdös-Rényi graph. Some works do acknowledge Gilbert's contribution, referring to this model as the Erdös-Rényi-Gilbert model, and it has also been called

the Bernoulli graph (Handcock, 2003a), Poisson model (Chakrabarti and Faloutsos, 2006), or classical random graph model (Kolaczyk, 2009).

A network model is defined by Kolaczyk (2009) as the collection

$$\{\mathbb{P}_\theta(G), G \in \mathcal{G}; \theta \in \Theta\} \tag{2.1}$$

where $\mathcal{G}$ is the set of all possible graphs and $\mathbb{P}_\theta$ is a probability distribution over $\mathcal{G}$ with parameter vector θ. Three common approaches to specifying the model, $\mathbb{P}_\theta(G)$, are discussed. The first is to restrict the set of graphs, $\mathcal{G}$, under consideration by specifying a set of features, such as a fixed number of nodes or edges. As in the specification of Erdös and Rényi, $\mathbb{P}_\theta$ is then specified as a uniform distribution over the resulting set of possible graphs. The next approach to specifying the model in (2.1) is to induce $\mathbb{P}_\theta$ through an algorithmic generating mechanism that simulates a graph. Random variables are often assigned to components of the generative process and probability distributions specified for the random variables. A limitation of this method from a statistical viewpoint is the often lack of a likelihood function for the entire graph. Although the generative algorithm may induce $\mathbb{P}_\theta$, in most instances it is prohibitively difficult to formulate and is rarely attempted. The last approach is to explicitly specify $\mathbb{P}_\theta$ by associating subgraph configurations and covariate information with graph topology of interest. This approach is taken by many statisticians in the field of network analysis. As a final note, the three approaches to model specification are not mutually exclusive, nor do they encompass all possibilities. For instance, the Erdös-Rényi-Gilbert model can be interpreted within all three categories, and some generative methods have only partial probability structures, thus, it is not clear how to take a generated graph and perform a probability analysis.

There are various possible categorizations of network analysis approaches. Those discussed here will be categorized as algorithmic construction or probabilistic modeling, where the distinguishing characteristic is the interest in a likelihood function. Methods under the heading of algorithmic construction involve the development of an algorithm-based graph generators that can quickly simulate a network which resembles an observed network of interest as much as possible with respect to features deemed to be important. They utilize the first two approaches

of network model specification and either a likelihood function cannot be identified or there is a lack of interest.

The goal of those methods under probabilistic modeling is "statistical model building" (Kolaczyk, 2009) and are defined by a likelihood. Probabilistic models for random graphs are specified by the third approach discussed above. These approaches allow for the estimation of parameters that provide a logical representation of the data and a method to evaluate and compare the fit of the competing models.

A network, or graph, is defined by a set of n nodes and m edges. Most of the discussion will focus on simple graphs, i.e., graphs with unweighted edges and no self-loops, with edges that can be directed or undirected. Graphs are observed at a single point in time, ignoring recent work on dynamic aspects of networks. Let V represent the set of vertices, or nodes and E the set of edges between pairs of nodes. A graph will be represented as G with edge values collected into $\mathbf{Y}$, an $n \times n$ adjacency matrix, with each entry Y_{ij} a binary random variable designating the presence, $Y_{ij} = 1$, or absence, $Y_{ij} = 0$, of an edge between nodes i and j. A realization of the graph will be represented as $\mathbf{y}$.

2.2 Graph Analysis: Algorithmic construction

The defining feature of the graph analysis techniques discussed in this section is the absence of a likelihood function for the graph, either from a lack of consideration or from an inability to discern its functional form. This work has been published largely in the computer science and statistical physics literature where the focus is on the ability to generate realistic graphs. Researchers in this area have condensed observed networks into common, seemingly important features where the goal of the proposed graph-generation algorithm is to quickly generate a graph with as many of the important features as possible. These algorithms may have parameters to be set and there is often probability involved in the formation and deletion of edges; however, given an observed network, these models do not have an ability to estimate values for the parameters, quantify uncertainty, or account for measurement error.

The motivation for developing algorithmic graph generators is to gain an understanding of the processes that lead to the formation of a network of interest (Leskovec et al., 2010)

because often the network to be analyzed is observed at a single point in time. Intuitively, if the algorithm results in a graph comparable to the network of interest, it is plausible that the observed graph arose as a result of operations similar to those performed by the algorithm. Understanding the graph formation procedure can lead to an ability to detect abnormalities in another observed network, allow one to compress a large network while preserving important features (Chakrabarti and Faloutsos, 2006), or to extrapolate and test out scenarios on graphs which cannot be observed (Leskovec et al., 2010), e.g., the Internet in five years.

Three features are commonly observed in realistic networks (Lancichinetti et al., 2008): a power law degree distribution, a small diameter, and clustering. Generators aim to emulate these three features exactly as they appear in a network of interest, in addition to as many other features as possible. For example, a recently proposed algorithm boasts the ability to simulate graphs which match realistic networks on 11 network characteristics (Goldenberg et al., 2010). Only the three generally agreed upon, important features will be described below.

The degree of a node is the number of edges incident, or connecting to, the node (Salter-Townshend et al., 2012). In an undirected graph, the degree of node i, $d(i)$, is found by summing over either the ith row or column of the adjacency matrix, $d(i) = Y_{+j} = \sum_{j=1}^{n} Y_{ij}$. The degree distribution is the collection of degrees for all nodes in the graph, $\{d(1), d(2), \ldots, d(n)\}$. Many networks of interest contain a few nodes with a large degree while a majority of the nodes have a small degree. In a social network, this is manifested as a few popular people, e.g., celebrities, with a lot of connections and ordinary people with fewer connections. In a graph of the Internet, there are a few websites to which many other sites link, e.g., Wikipedia, Google, while the vast majority have substantially fewer. This phenomena suggests the degree distribution is heavily right skewed, or, as is more commonly described, follows a power law with probability density function of the form

$$p(d(x)) = A \times d(x)^{-\gamma} \tag{2.2}$$

where $A > 0$ is a normalizing constant and $\gamma > 1$ is the power law exponent. Networks with this property are referred it as scale-free graphs (Ben-Avraham et al., 2003). The value of γ is often used as a metric to determine how well the algorithm is able to replicate the degree

distribution of the observed network (Bar et al., 2007). However, estimation of the exponent is not straightforward nor is its computation consistent. Chakrabarti and Faloutsos (2006) list seven of the more commonly used methods.

The second important feature is a measure of graph connectedness. Distance between two nodes can be defined as the number of edges on the shortest path between them. If no such path exists, the distance is defined to be infinity. A graph is connected if all distances are finite and unconnected otherwise. The diameter of a graph is the maximum distance between all pairs of nodes (Gross and Yellen, 2006). For an unconnected graph, either the maximally connected subgraph or the effective diameter can be considered (Salter-Townshend et al., 2012), where the effective diameter is the minimum number of edges between some percentage of nodes (Chakrabarti and Faloutsos, 2006). The diameter in empirical networks has been found to be quite small, especially compared to the size of the network, resulting in the "small-world" effect. For example, Watts and Strogatz (1998) examined a graph with 225,226 movie actors as nodes with an edge between actors in the same film. Using the maximally connected subgraph, the diameter was found to be 3.65, making Kevin Bacon seem less impressive. In the same work, the US power grid represented as 4,941 nodes is found to have a diameter of only 18.7.

Clustering is the final important feature for graph generators to replicate. This phenomena refers to the large number of triangles in an empirical network and is also referred to as transitivity. In a social network, the interpretation of transitivity is that two individuals are more likely to be friends if they share a common friend. Newman et al. (2002) claim the probability of edge formation between two nodes is several orders of magnitude greater if those nodes have a distance of two between them. The amount of clustering is represented by the clustering coefficient, C, which quantifies the proportion of the connected sets of three nodes, or triples, which are closed and thus form a triangle,

$$C = \frac{3 \times \text{Number of triangles}}{\text{Number of connected triples}}. \tag{2.3}$$

Empirical networks have been found to have a larger value of a clustering coefficient than if edges formed independently and at random. The movie actor example (Watts and Strogatz, 1998) has a clustering coefficient of 0.79, so 79% of connected triples are triangles. The authors

contrast this with a single graph of the same number of nodes and edges generated by placing the edges at random which resulted in a clustering coefficient of 0.00027.

Some limitations to the graph generation algorithmic approach to network analysis have been identified in the literature. First, although the algorithms attempt to recreate as many features of the empirical networks as possible, there is no consideration if these features are sufficient or necessary descriptions of network structure (Fienberg, 2012). The important features are chosen because they appear frequently in observed graphs, although recent analysis suggests some features are not as ubiquitous as previously believed (Goldenberg et al., 2010). In fact, Bar et al. (2007) suggest that the power law degree distribution demonstrated in the Autonomous Systems (AS) network, a crucial component of Internet connectivity, may be a consequence of the manner in which the data were collected. Descriptions used to summarize the algorithms do not explore the full parameter space (Goldenberg et al., 2010), and parameters are given as point estimates without any quantification of uncertainty. Thus, it is possible that the summary quantities of the realistic networks are highly inaccurate (Fienberg, 2012).

Statistical methods for estimating the model parameters of observed data are also lacking (Fienberg, 2012). When statistical methods are used, they are often applied incorrectly or in violation of assumptions. As an example, to estimate the power law exponent for a degree distribution, γ in (2.2), the distribution is plotted on a log-log scale and the slope is obtained either through ordinary least squares or visual inspection. This approach is used even in the presence of strong non-linearity (Goldenberg et al., 2010). Further, measurement error or other potential biases in the data are not considered. Bar et al. (2007) suggest that up to 50% of the edges in the AS network are not observed; however, even with this acknowledgement, the authors claim they "cannot model data that are unknown."

Despite the statistical limitations of graph generating algorithms, they have been given a lot of attention in the statistical physics and computer science literature. In 2010, Kolaczyk (2010) speculated that at least 2/3 of the published work on network analysis focused on descriptive methods and as Goldenberg et al. (2010) stated, "Alternative graph generation mechanisms appear [in the literature] every day." A few historically important and illustrative examples of

the graph generating algorithms will be presented below. Many of the algorithms not included were formulated as slight variations of those discussed.

2.2.1 Random Graph Models

Random graph models (RGMs) are those for which the set of possible graphs $\mathcal{G}$ has been defined and equal probability is place on each graph, $G \in \mathcal{G}$ (Kolaczyk, 2009). These are formulated according to the first common approach of defining the network model, (2.1). A RGM is completely determined by identifying the set of plausible graphs, $\mathcal{G}$, which can be accomplished in two ways: by explicitly stating the set or by determining the possible graphs that could arise from a graph generating algorithm.

In the context of a RGM, the Erdös-Rényi-Gilbert model is often referred to as the "classical random graph model" or just *the* random graph. As mentioned previously, the Erdös-Rényi-Gilbert model can be cast as all three types of graph model formulation for (2.1). In addition, the specifications from Erdös and Rényi and from Gilbert can be used to demonstrate the two ways of defining the set $\mathcal{G}$. Under the specification of Erdös and Rényi, the set $\mathcal{G}$ contains graphs with a fixed number of n nodes and m edges. Gilbert's specification can be considered a graph generation scheme, a phenomena referred to by Goldenberg et al. (2010) as "pseudo-dynamic," where models that were originally proposed to describe a single, static network can be interpreted as a generative algorithm. For Gilbert's specification of the Erdös-Rényi-Gilbert model, the process begins with n disconnected nodes. At each iteration of the algorithm, a pair of nodes is considered and an edge is added between them with probability $p = m/\binom{n}{2}$, independent of all previous iterations. This continues until all pairs of nodes are considered. The set of nodes remain fixed and once an edge is added, there is no mechanism to remove it. The set $\mathcal{G}$ resulting from each specification will be equivalent as $n \to \infty$.

One reason the Erdös-Rényi-Gilbert model has received so much attention is that many of its properties can be calculated exactly (Newman et al., 2002), specifically the identification of a "phase change" (Fienberg, 2012) or "phase transition" (Chakrabarti and Faloutsos, 2006), which occurs at $\lambda = pn = 1$. When $\lambda < 1$, graphs contain small, disconnected groups of edges. The phase associated with $\lambda > 1$ is characterized by one giant component (Fienberg, 2012),

which occurs when a majority of the nodes are highly connected (Kolaczyk, 2009). This phase is more commonly observed in empirical networks, and Newman et al. (2002) add the existence of a giant component to the list of features of a realistic network. A direct result of the giant component is the small-world property. However, Erdös-Rényi-Gilbert graphs with $\lambda > 1$ fail to reproduce the other two important features (Chakrabarti and Faloutsos, 2006). The degree distribution is Binomial and approaches a Poisson as $n \to \infty$ and therefore, the graphs resulting from the generative method are not scale-free. In addition, because edges form independently, so do triangles, and thus the graphs lack the desired clustering. Often, the Erdös-Rényi-Gilbert model is used as a "straw-man" model for newly-proposed algorithms.

In order to address some of the shortcomings of the classical RGM, one proposed solution is to further restrict $\mathcal{G}$ to only graphs that contain an important, omitted feature. Graph generation algorithms of this type are called generalized RGMs. Modifications to the set $\mathcal{G}$ are intended to produce graphs with features such as small clusters of highly connected nodes, more realized triangles (Fienberg, 2012), or most commonly, a specified degree distribution (Aiello et al., 2001). For this last type of restrictions, the number of nodes and the degree distribution of the graph are fixed, which only allows for a specific number of vertices. Thus, the set of possible graphs for this type of generalized random graph model is a subset of those allowed under the Erdös-Rényi-Gilbert model, given Erdös and Rényi's specification.

The method to simulate a graph with a specified number of nodes and a degree distribution is as follows. Begin with a graph of n unconnected nodes. Each node is randomly assigned a degree. Nodes are then joined until none of the nodes have any extra degrees. The standard algorithms developed to perform this last step are the matching algorithm and switching algorithm (Kolaczyk, 2009).

Similar to a classical RGM, mathematical properties can be solved in the limit of large n for a generalized RGM that fixes the degree distribution. Specifically, if the degree distribution is defined as a power law, (2.2), the existence and size of a giant component can be determined as a function of A and the power law exponent, γ. The diameter and average path length of these generalized random graphs can also be determined (Aiello et al., 2001). More generally, if the specified degree distribution is not a power law, the emergence of the giant component

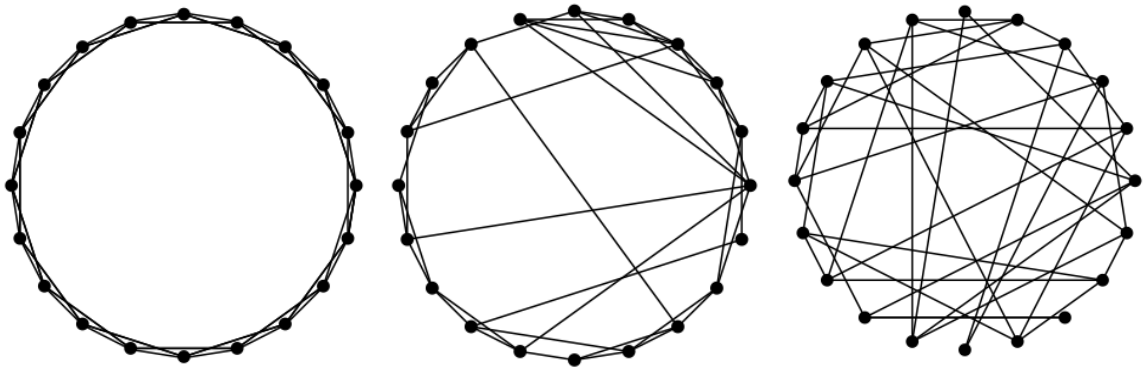
can be computed based on the first two moments of the degree distribution, and its size can be determined from the number of nodes.

The criticism of this type of generalized RGM is that the resulting graphs only match the degree distribution, and if the giant component exists, then the small world property as well. These graphs often do not contain the high level of transitivity, the third important feature for graph algorithms to produce. In addition, this model cannot distinguish between two graphs which have the same degree distribution but with structure that differs according to other metrics (Krivitsky et al., 2009).

2.2.2 Small World Models

Network analysis began with the classical RGM and a goal of understanding its properties. As the number and availability of observed networks increased, the limitations of the classical random graph model as an adequate representation of reality became more clear. This realization prompted what Kolaczyk (2009) refers to as a significant historical shift in the approach to network analysis. The move was away from a theoretical understanding of the random graph model and to the creation of models designed to explicitly generate a graph with features of interest. Clearly, the generalized RGM could be considered of this type with its ability to recreate, for example, a specified degree distribution exactly. A seminal work that helped to spur this change to graph generation is the introduction of the small-world model (Watts and Strogatz, 1998).

A small-world model is highly connected and transitive (Goldenberg et al., 2010), with a small diameter and large clustering coefficient (2.3). The combination of these two features is not possible with a RGM because as the number of nodes increases the diameter also increases, and transitivity is inversely related to the number of nodes (Kolaczyk, 2009). In the development of the small-world model, Watts and Strogatz (1998) envisioned a spectrum of randomness for networks. At one extreme is a regular graph with no randomness. The example regular graph used in this work consists of n nodes equally spaced on the circumference of a circle where each node is connected to the closest k nodes. This implies an edge between node i and the closest $k/2$ nodes clockwise on the circle and $k/2$ in the counter-clockwise direction



(a) Regular, $p = 0$ (b) Small World, $p = 0.2$ (c) Random, $p = 1$

Figure 2.1 Demonstration of the Small-World model of Watts and Strogatz (1998). The graphs increase in randomness from right to left.

from node i. An example with $n = 20$ and $k = 4$ is shown in the far left plot of Figure 2.1. Regular graphs have a high value of the clustering coefficient, (2.3). At the other end of the randomness spectrum is a random graph. A random graph contains the same number of edges, $m = kn$, but each edge connects two nodes chosen independently and at random, while excluding the possibility of self-loops and multi-edges. An example of a random graph with $n = 20$ and $m = 4 \times 20 = 80$ is shown on the far right of Figure 2.1. Random graphs have a small diameter, i.e., exhibit small-world behavior. The Small-World model, also known as the Watts-Strogatz Model, falls between the regular and random graphs on the randomness spectrum, while retaining both features of high clustering from a regular graph and the small-world property from a random graph.

The small-world model can also be cast as a generation mechanism and thus classifies as a "pseudo-dynamic" model (Goldenberg et al., 2010). The process of generating a small-world model begins with a regular graph of n nodes and k connections. Each edge is considered in turn and has a fixed, independent probability p of being rewired. If an edge is chosen to be rewired, one end of the edge is relocated to a different node, chosen uniformly from the remaining $n - 2$ nodes. Again, self-loops and multi-edges are disallowed, thus the number of edges remains constant at $m = kn$. The middle plot of Figure 2.1 shows a small-world network

with probability to rewire, $p = 0.2$. The parameter p determines the location of the graph on the randomness spectrum. In the extremes, $p = 0$ implies that no edges are rewired, resulting in the regular graph and $p = 1$ implies that all edges are rewired, and thus a random graph is simulated.

The disadvantage of a graph produced from the small-world model is that the graph will not have a power law degree distribution. A regular graph has a degenerate degree distribution at degree k, and the degree distribution of a random graph is Binomial. A graph generated from the small-world model will have a degree distribution somewhere between the two extremes; however, because the value of p is often chosen to be small, the degree distribution more often resembles the degenerate distribution of a regular graph. In addition, formal statistical methods do not exist to assess the fit of the model to empirical networks (Goldenberg et al., 2010). Theoretical properties of graphs simulated from the small world model are not easily determined and are described by Kolaczyk (2009) as an "open problem".

2.2.3 Preferential Attachment

A preferential attachment model is categorized as a "network growth model" (Kolaczyk, 2009) because models of this type are designed to model the evolution of a network over time. The basis of the modern varieties of preferential attachment models was developed by Barabási and Albert (1999) specifically to model the expansion of the World Wide Web. The authors were motivated by the observation that new webpages were more likely to form links to the more popular, currently existing pages than the less popular ones. The rationale behind the approach has been referred to as "The rich get richer" or cumulative advantage (Chakrabarti and Faloutsos, 2006) and is related to Zipf's Law and the Chinese Restaurant Process (Vivar and Banks, 2012).

In contrast to the previously discussed network generating approaches, a preferential attachment model allows for the addition of nodes. The process of generating a network begins with n_0 nodes and m_0 edges. At each iteration one new node is added to the network and connected to $q < n_0$ existing nodes. The probability a new node connects to an existing node v is proportional to its degree. If $d(v)$ represents the degree of node v, the probability the newly

added node will connect to node v is

$$p_v = \frac{d(v)}{\sum_i d(i)}. \tag{2.4}$$

After t iterations the graph will contain $n_0 + t$ nodes and $m_0 + qt$ edges. An important property of graphs produced from the preferential attachment model is a power law degree distribution, i.e., a scale-free graph. As the number of iterations grows, the power law exponent γ in (2.2) approaches 3 (Kolaczyk, 2009) regardless of the number of nodes added at each iteration, q (Chakrabarti and Faloutsos, 2006). The resulting graph also exhibits a small-world behavior. Asymptotic bounds have been determined for the diameter of a preferential attachment graph that relates to the number of nodes logarithmically (Leskovec et al., 2007).

The lack of flexibility of networks generated from the original preferential attachment method has been criticized. First, the graphs may exhibit small-world behavior, but the model does not include a parameter to control it (Vivar and Banks, 2012). Another feature not controlled by the model is the power law exponent that always approaches $\gamma = 3$. Although the shape of the degree distribution, particularly the tails, may change as new nodes are added, the average degree remains constant at q, while empirical evidence suggests the average degree should increase as the network grows (Chakrabarti and Faloutsos, 2006). The model is unable to produce networks with a dense core because edges are always added q at a time (Bar et al., 2007), and the method is unable to produce graphs with several connected components or isolated nodes (Chakrabarti and Faloutsos, 2006). Finally, the growth of the diameter as the number of nodes increases does not match reality as recent evidence suggests the diameter actually shrinks as the network grows (Leskovec et al., 2010).

When a deficiency of the original preferential attachment model is presented, it is often followed with a modified version that addresses the stated inadequacy. The simplicity of the original approach (Barabási and Albert, 1999) has also contributed to the many existing extensions. Chakrabarti and Faloutsos (2006) detail ten of these extension and the inadequacy of the original preferential attachment model that is addressed. As an example, an initial attractiveness model allows for a more general power law by adding a parameter to the edge

connectivity probability in (2.4) so that it becomes

$$p_v = \frac{d(v) + A}{\sum_i [d(i) + A]}$$

with the resulting power law exponent $\gamma(A) = 2 + \frac{A}{q}$. The power law is now a function of the parameter A. The forest fire model is an example of a more elaborate variation of the original preferential attachment model (Leskovec et al., 2007). Networks generated from the forest fire model are scale free, have a decreasing diameter, an increasing average degree, are directed, and allows for community structure. Two additional parameters are used in this model: a forward burning probability, p_{fb}, and a backward burning ratio, r_{bb}. At each iteration, a new node is added to the graph and edges form according to the following steps:

1. Choose an "ambassador node," w, uniformly at random from the existing nodes of the graph. Form a link between the new node to w.

2. Draw a random number, n_1, from the binomial distribution with mean $(1 - p_{fb})^{-1}$.

3. Choose n of the currently existing edges of node w. Select edges that are directed to node w with probability r_{bb} times less than edges that are directed away from node w. Let $w_1, w_2, \ldots, w_n$ represent the nodes at the other ends of the edges selected.

4. Connect the newly added node with a directed edge to $w_1, w_2, \ldots, w_n$.

5. Repeat steps 2 and 3 recursively for each of the $w_1, w_2, \ldots, w_n$

Extensions to the forest fire method have also been developed to allow for isolated nodes, or orphans, or to choose multiple ambassador nodes.

Preferential attachment models are an example of how many of the algorithmic graph generators are applied to realized networks, and the forest fire model is an example of how complicated the algorithms can become. Simulated networks are mostly used to compare characteristics from a realized network of interest. The goal is for the simulated network to match the realized graph on characteristics studied. Metrics and statistics have also been developed to test the resemblance of the simulated graph to the network of interest. Often

these values relate back to the three main features of a network: power law degree distribution, small diameter, and clustering. Little effort has been made to estimate the parameters of the model given an observed network.

2.3 Graph Analysis: Probabilistic modeling

In contrast to the algorithmic graph generators of the previous section, models described in this section can be described with a likelihood function. Therefore, a statistical model can be constructed for an entire graph with a joint distribution, and statistical inference can be conducted. Controlling parameters can be estimated, the probability of a realized graph can be determined, and the fit of the model can be assessed. It should be noted that although it is possible to generate graphs from models of this type, these models go beyond generation.

Two broad classes of probabilistic modeling of random graphs will be discussed in detail. The first class are exponential random graph models (ERGMs), which specify a joint distribution for the collection of edge variables with the goal of describing global network features through interactions of local edge configurations. Conversely, models categorized as latent variable models (LVMs) will focus on the interpretation of properties of the individual nodes of the graph (Hunter et al., 2012). This category encompasses a wider range of graph models that are hierarchical in nature. Conditional distributions for edges variables specify the model and are considered to be independent given some latent variable, such as block membership or position within a social space.

In reference to social networks, Snijders (2007) described ERGMs and a subclass of LVMs, the latent space models, as the two competing models for probabilistic modeling and statistical analysis of networks. Although specific applications may be well-suited for one approach or the other, there are networks where both would be applicable. Each method presents advantages and challenges, a brief summary of which is presented in Table 2.1. Models that integrate the advantages of both approaches while minimizing the difficulties are referred to as "the next generation" of models (Snijders, 2007).

Table 2.1 Comparison of the strengths and weaknesses of the two categories of probabilistic modeling: exponential random graph models (ERGMs) and latent variable models (LVMs).

	Advantage	Disadvantage
ERGM	⋆ Scientific justification of statistics ⋆ Can incorporate many network features	⋆ Can become degenerate ⋆ Likelihood is intractable ⋆ Cannot account for unobserved structure
LVM	⋆ Computationally tractable ⋆ Not degenerate ⋆ Can handle data that is missing at random	⋆ Cannot model dependencies between edges ⋆ Cannot account for transitivity

2.3.1 Exponential Random Graph Models

The exponential random graph model (ERGM) is a widely-used and extensively-studied class of models under the category of probabilistic modeling. This class arose from collaborations between sociology, psychology, and statistics and was originally developed to model social networks. Unlike many of the other analysis techniques discussed, ERGMs have been applied to networks in areas other than that for which it was originally intended (e.g., Groendyke et al., 2012; Simpson et al., 2011). Its popularity can be partially attributed to the model's ability to represent graph topology while also allowing for complex dependencies.

2.3.1.1 Introduction

An ERGM is specified as a joint distribution for that adjacency matrix, $\mathbf{Y}$, in exponential family form (Kolaczyk, 2009). The exponential family was chosen because the sufficient statistics are explicitly tied to parameters and are equal to their expected values (Holland and Leinhardt, 1981). Let $\Omega \equiv \{\mathbf{y} : \Pr(\mathbf{Y} = \mathbf{y}) > 0\}$ be the support of the joint distribution. A general functional form of the joint distribution is

$$\Pr(\mathbf{Y} = \mathbf{y}) = \frac{1}{\kappa} \exp\left\{ \sum_{T \subseteq C} \theta_T g_T(\mathbf{y}) \right\} \tag{2.5}$$

where

- C is the set of all paris of nodes between which an edge could form; most often all pairs of nodes, so $|C| = \binom{n}{2}$
- $T \subseteq C$ is a subset of the possible edges, often called a configuration
- θ_T is a parameter corresponding to configuration T
- $g_T(\mathbf{y}) = \prod_{(i,j)\in T} y_{ij}$ is a network statistic, which is equal to 1 when configuration T occurs in $\mathbf{y}$
- $\kappa = \sum_{\mathbf{Y}} \exp\left\{\sum_{T\subseteq C} \theta_T g_T(\mathbf{y})\right\}$ is the normalizing constant.

The summation in the exponent of (2.5) is referred to as the negpotential function

$$Q(\mathbf{Y}) = \sum_{T\subseteq C} \theta_T g_T(\mathbf{y}). \tag{2.6}$$

Defining the negpotential function (2.6) defines the joint distribution, up to a constant. Specifying a particular set of parameters, θ_T, or equivalently network statistics, $g_T(\mathbf{y})$, to be included in the negpotential specifies an ERGM. Often these statistics are counts of subgraph features, such as the number of edges, number of triangles, or number of edges in a particular block, and the corresponding parameters represent density, transitivity, and block effect, respectively.

Early model development for ERGMs involved the introduction of parameters or statistics to be included in the negpotential function (2.6). Four main phases of ERGM development will be discussed. First, the proposal of a model of form (2.5) for network analysis. Next, the relaxing of an independence assumption to allow for more complex dependence structures, followed by the proposal of parameters not motivated by dependence. The most recent development involves the introduction of parameters designed specifically to address a common issue with the application of an ERGM to realized networks, that of model degeneracy.

Many studies point to the seminal work of Holland and Leinhardt (1981) as the introduction of ERGMs. This work proposed the p_1 model, a log-linear model applied to the dyads of a directed social network. A dyad is defined as a pair of nodes and the possible ties between them, which in an unweighted, directed graph could be 0, 1, or 2. Prior network analysis was descriptive only, focusing on aspects such as the degree distribution or the distribution of nodal

attributes. In contrast, Holland and Leinhardt (1981) were able to stochastically model the patterns of relationships. The p_1 model allows for a simultaneous estimation of a parameter representing reciprocity, or a tendency for a edge to be reciprocated, and a parameter for differential attractiveness, which occurs when a node attracts a comparatively large number of edges. The authors argued both reciprocity and differential attractiveness were common in realized social networks.

Although the p_1 model was the first to be able to estimate meaningful parameters, a disadvantage is that the dyads are modeled to be independent. This assumption leads to a likelihood that is a product of probabilities for each dyad, which was necessary in order to estimate the parameters with the existing statistical techniques of the time. Relaxing this assumption was identified by Holland and Leinhardt (1981) to be important, yet difficult. The next ERGM development was introduced by Frank and Strauss (1986), who were able to incorporate a more general definition of dependence by adapting methods developed for spatial statistics (Besag, 1974) to social networks. Because of the ability to model a general, complex dependence structure, it is most common for Frank and Strauss (1986) to be cited as the origin of ERGMs.

To illustrate the complex dependence structure of a network, Frank and Strauss (1986) introduced the dependence graph for social network analysis. Each node in the dependence graph corresponds to a potential edge in the original graph where a connection in the dependence graph indicates the corresponding random variables are conditionally dependent. An important dependence structure is incidence, or Markovian, in which two edges are conditionally dependent if they share a common node, and graphs with this dependence structure are referred to as Markov graphs Frank and Strauss (1986). The dependence graph corresponding to a Markov graph does not have edges between disjoint sets of nodes. Stated another way, let $\{i, j\}$ and $\{m, n\}$ represent two potential edges in the original, Markov graph and thus two nodes in the dependence graph. If $\{i, j\} \cap \{m, n\} = \emptyset$ the two edges are conditionally independent, and there will not be an edge between these nodes in the dependence graph. The set of random variables on which a particular random variable is conditionally dependent, i.e., linked in the dependence graph, will be called its neighborhood. For a Markov graph the neighborhood of edge $\{i, j\}$ is $\{\{r, s\} : \{i, j\} \cap \{r, s\} \neq \emptyset\}$.

Closely related to the idea of neighborhoods is the Hammersly-Clifford Theorem. This is an important theorem that has been stated and proven in various forms and in multiple references. (For the original see (Clifford, 1990); for one similar to what is used here see (Cressie, 1993, p. 417)) First, assume the network contains a finite number m of possible edges. Let the support of the joint distribution of $\mathbf{Y}$ be designated as Ω and assume there exists a $\mathbf{y}^* \in \Omega$ such the that joint probability distribution is positive, $\Pr(\mathbf{Y} = \mathbf{y}^*) > 0$. Besag (1974) showed that for $\mathbf{y}^* = 0$ the negpotential function can be expanded uniquely over Ω as a summation over configurations of random variables, i.e., that the negpotential takes the form shown in (2.6). This result was later generalized for any $\mathbf{y}^* \in \Omega$ (Kaiser and Cressie, 2000). The Hammersly-Clifford Theorem states that the parameter, θ_T, will be non-zero only if the random variables in the corresponding configuration T form a clique, where a clique is a single random variable or a set of random variable such that every pair within the set is pairwise mutually conditionally dependent, given the rest of the graph. This theorem implies the dependence structure affects which parameters in (2.6) will be non-zero.

As an example, the non-zero parameters for a Markov graph will correspond to triangles and k-stars. A k-star configuration results from k edges which all share a common node, with order ranging from $k = 1, \ldots, n-1$. Note that a 1-star is just an edge in the graph. In practice, all order stars are rarely considered because of the large number of parameters this model would include; however, the dependence structure is still incidence. A common model used as an example is the triad model, which consists of a term for 1-stars, i.e., edges, 2-stars, and triangles. The negpotential function of the triad model is

$$Q(\mathbf{Y}) = \rho g_1(\mathbf{y}) + \sigma g_2(\mathbf{y}) + \tau g_3(\mathbf{y})$$

where ρ is a density parameter, σ a parameter for clustering, τ is a parameter that represents transitivity, and the network statistics count the corresponding number of configurations in $\mathbf{y}$ (e.g., $g_2(\mathbf{y})$ counts the number of 2-stars in the network).

The parameters, θ_T, included in the negpotential function of a Markov graph are the result of the incidence dependence structure and the Hammersly-Clifford Theorem. The next refinement in ERGM specification began with Wasserman and Pattison (1996) who proposed parameters

and statistics motivated by the graph topology they represent, rather than the dependence structure. These models were named p^* in honor of the p_1 model (Holland and Leinhardt, 1981), which also included parameters motivated by empirical observation of networks. Four tables of possible parameters are introduced with the p^* model with a statement that those listed are just a subset of the possibilities.

As an application of the converse of the Hammersly-Clifford theorem, the terms included in the negpotential function of a p^* model can be used to determine neighborhoods of edges. Thus, a conditional dependence structure is induced by the choice of terms included in the moel. Five methods are suggested for how to choose the parameters of a p^* model (Wasserman and Pattison, 1996), one of which is to consider the induced conditional dependence between possible edges. Although, even if the conditional dependence is not a modeling consideration, there is still an induced conditional dependence structure. For some example p^* models, Wasserman and Pattison (1996) describe the induced conditional dependencies, but for others admit that identifying the sets of conditionally dependent edges is not immediate and that some models induce "arbitrary complexity." Further, varying the parameters in the negpotential can lead to changes in the implied dependence structure. More recently, Goodreau (2007) pared down the list of parameter-choosing methods to two approaches: considering the implied dependence or the combination the of parameters that fit the empirical network best.

The extension to include arbitrary statistics in the negpotential function (2.6) provides a straightforward approach to incorporating exogenous information in an ERGM. The parameters discussed up to this point have been endogenous, or functions of the graph itself (e.g., density or transitivity). Exogenous attributes do not depend on the structure of the graph and can be incorporated at the level of the individual nodes, edges, or as symmetric functions. A main effect term is an example of how to include an exogenous attribute of a node. This parameter varies with the covariate value of the node (Goodreau, 2007). A similar example for pairs of nodes is assortative mixing which attempts to capture the increased probability of edges to form between nodes within the same attribute class (Goodreau et al., 2008). When this effect is uniform across attribute classes, it is referred to as uniform homophily. For example, consider a friendship network between grade-school aged children. A uniform homophily term

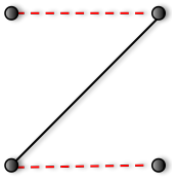


Figure 2.2 Configuration of edges which leads to partial conditional dependence.

could represent the higher probability of friendships to form between children within the same grade. Differential homophily allows for a different parameter to describe this effect within the different attribute classes. For the grade-school example, a differential homophily term could be used to describe how friendships are more likely to form between two nodes of the same gender *and* how females tend to make more friends than males (Hunter et al., 2008b). As a final example, if the covariate is continuous, such as age, the absolute value of the difference can be included as a statistic, and the corresponding parameter would allow the probability to vary monotonically with the value of the absolute difference (Goodreau, 2007). Attribute-based parameters do not affect the dependence structure and thus, including only these parameters leads to a dyad-independence model.

An approach that considers the effect of the dependence structure and higher-order statistics are the two extensions to ERGMs termed "neighborhood-based" models (Pattison and Robins, 2002). The first extension uses "social settings," or groupings of nodes, typically based on nodal attribute values. Multiple social settings can be defined for a model, and they may be overlapping. Two edges are considered conditionally independent if they do not occupy any common setting, although they are not necessarily conditionally dependent if they do. The purpose of this approach is to limit the number of conditionally dependent random variables, or the size of the neighborhoods. The second extension includes a parameter to account for configurations of at least four nodes such that every pair of edges are part of a path of length three. An example of such a configuration is shown in Figure 2.2. The dependence structure induced by this form of statistics is termed partial conditional dependence. Two random variables are defined to be partially conditionally dependent, if their dependence status is

determined by the state of a third random variable. For the example shown in Figure 2.2, the two dashed, red lines will lie on a path of length three and therefore be conditionally dependent only if the solid, black edge is realized.

Other extensions to the original ERGM that will not be discussed in detail here include an extension to networks with multiple relations (Pattison and Wasserman, 1999), networks with values on the edges (Robins et al., 1999), and including the attributes into the dependence graph to predict node-level attributes given the graph topology (Robins et al., 2001). The final, significant contribution to the development of model parameters for an ERGM to be discussed is the proposal of the alternating or geometrically weighted statistics (Snijders et al., 2006; Robins et al., 2007). Full consideration of this new specifications will be postponed until the section on degeneracy, Section 2.3.1.3, as the terms were developed to combat this specific issue.

Although an ERGM is typically expressed as a joint distribution, it can also be expressed as a conditional log-odds, which allows for a more natural parameter interpretation. Consider the random variable, Y_{ij}, representing a potential edge between nodes i and j. The conditional log-odds for Y_{ij} implied by the joint specification of an ERGM (2.5) is

$$\text{logit}[\Pr(Y_{ij} = 1 | \mathbf{Y}^c_{ij} = \mathbf{y}^c_{ij})] = \sum_{T \subseteq C} \theta_T \boldsymbol{\delta}_g(\mathbf{y})_{ij} \tag{2.7}$$

where $\mathbf{Y}^c_{ij}$ represents all edges in the network other than edge $\{i, j\}$ and $\boldsymbol{\delta}_g(\mathbf{y})_{ij}$ is the vector of change statistics. A chance statistic is

$$\delta_g(\mathbf{y})_{ij} = g_T(\mathbf{y}^+_{ij}) - g_T(\mathbf{y}^-_{ij})$$

where $g_T(\mathbf{y}^+_{ij})$ is the value of the network statistic if edge $\{i, j\}$ is present while the rest of the graph remains unchanged and $g_T(\mathbf{y}^-_{ij})$ is the analogous value when edge $\{i, j\}$ is absent. The change statistic, $\delta_g(\mathbf{y})_{ij}$, represents the effect on the network statistic $g_T(\mathbf{y})$ if random variable Y_{ij} is changed from 0 to 1 and all other random variables remain constant. As an example, if $g_T(\mathbf{y})$ represents the density of the graph, the change statistic is always 1 because the addition of an edge will always increase the density by 1.

A parameter, θ_T, of an ERGM is interpreted as the increase in conditional log-odds of a network as a result of a unit increase in the corresponding statistic. The conditional probability

for Y_{ij} depends on the rest of the graph, $\mathbf{Y}^c_{ij}$, only through the change statistic. If the model includes a homogeneity assumption, i.e., all isomorphic graphs are equivalent, the parameter values indicate the type of edges which are most probable. Individual parameters interpretation is heavily dependent upon the other terms included in the model. For example, if the model includes two k-star statistics for $k_1 < k_2$, then the interpretation of the parameter for k_2 is the effect on the conditional log-odds of the network due to k_2-stars adjusted for the number of k_1 stars (Kolaczyk, 2009). Although this is partially due to the fact that there are $\binom{k_2}{k_1}$ k_1-stars within a k_2-star, a confounding of interpretation occurs even if the statistics are not nested.

2.3.1.2 Estimation and goodness-of-fit

Exact maximum likelihood estimation (MLE) has never been a viable option for ERGMs. Computing the MLE would require repeated evaluation of a normalizing constant that involves a sum over all possible graphs. As the number of possible graphs grow super-exponentially with the number of nodes, $|\mathcal{G}| = 2^{\binom{n}{2}}$ this is computationally intractable even for trivially small networks (Rinaldo et al., 2009). For example, the number of simple graphs that can be formed from only 10 nodes is over 35 trillion. Thus, since the inception of ERGM, approximation methods have been devised to estimate the parameters. Unfortunately, model proposals have outpaced the estimation methods. The discussion below will focus on the more commonly used techniques for estimating the parameters of an ERGM: maximum pseudo-likelihood, two approximate maximum likelihood estimation techniques based on Monte Carlo simulations, and some recently proposed methods in Bayesian estimation.

Maximum pseudo-likelihood estimation (MPLE) was developed in the context of lattice (Besag, 1974) and non-lattice (Besag, 1975) data for applications of spatial statistics. It was applied to social networks and ERGMs by Frank and Strauss (1986) and Strauss and Ikeda (1990). The pseudo-likelihood (PL) function is an approximation to the joint distribution and is computed as the product of the full conditional distributions. For ERGMs, this is product of the conditional distributions shown in (2.7). The problematic normalizing constant cancels out in the conditional distributions, and thus the PL function is always available in closed form.

A reason for the widespread use of MPLE was the ease and speed at which the PL can be maximized. Strauss and Ikeda (1990) proved that the MPLE is equivalent to the maximum likelihood of a logistic regression where the data plays the role of both the dependent and independent variables. Using standard statistical software the PL can therefore be maximized through an iteratively reweighed Gauss-Newton least squares procedure. However, the PL function is not the true likelihood for any model (Geyer and Thompson, 1992), the two functions are simply computationally identical. Standard errors reported from the logistic regression fit are not applicable to the PL maximizers because logistic regression presumes the independent variables are fixed and dependent variables are random. For the PL scenario both are random since they both are copies of the data.

Although the PL approach is fast and provides an intuitive solution to the intractable normalizing constant, its use has been heavily criticized in the network analysis literature. The most common complaints are that the PL overestimates dependence and structural effects (Lubbers and Snijders, 2007), underestimates standard errors, and performs poorly in practice. These issues are exacerbated when the dependence between random variables is strong. Alternatively, if random variables are independent, maximizing the PL is equivalent to maximizing the likelihood. In a case study using DNA fingerprinting, Geyer and Thompson (1992) compared their MLE approach to MPLE and found that MPLE estimated much higher values of a dependence parameter, produced unreasonable probabilities, and overall provided a very bad fit to the data. Robins et al. (2007) compared standard errors between maximum likelihood (ML) and PL and found that on average PL standard errors were smaller, but could differ from ML by three to four times in either direction. Other identified issues with the PL method are that it is not admissible for a squared error loss function due to the fact that it is not a function of complete sufficient statistics (Snijders, 2002), it can produce infinite values even if the function converges (Handcock, 2003a), and that, due to its approximate nature, it can fail to indicate when the model has become degenerate (Robins et al., 2007). Lastly, the properties of the MPLE, specifically within the context of ERGMs, are not well understood and there is no asymptotic theory to base confidence intervals and hypothesis tests (Kolaczyk, 2009).

Besag (2001) suggests that the MPLE is not likely to perform well for ERGMs unless the dependence between edge variables is weak. It has been argued that ERGMs are more global than local (Caimo and Friel, 2011; Hoff et al., 2002) and the PL considers only local information (Handcock, 2003a). The PL function does not take into consideration the parameter space or normalizing constant. Thus, if this space is constrained or if the normalizing constant is an important aspect of the model, then PL is likely not to be an adequate estimation technique. Besag (1992) describes MPLE as "a simple tool from another era" that will eventually become obsolete and recommends the use of approaches that utilize simulation methods due to the increase in computing capabilities (Besag, 2001).

In response to the poor performance of the PL, two Monte Carlo approaches were developed to approximate the parameters of an ERGM. The first is Markov Chain Monte Carlo-Maximum Likelihood Estimator (MCMC-MLE) which is based on a stochastic approximation of the log-likelihood and a maximization of the approximation. The approach is based on the work presented in Geyer and Thompson (1992), and Hunter and Handcock (2006) were the first to apply it to ERGMs. The second approach approximates the MLE through a stochastic approximation to the moment equation (Snijders, 2002). It is based on the Robbins-Monro algorithm, a stochastic version of the Newton-Raphson algorithm. Both methods assume that it is possible to obtain a sample of graphs from a specified ERGM. Note that this is most directly accomplished through a Gibbs sampler using the full conditional of the log-odds shown in (2.7) to update random variables in turn. Other methods that update groups of variables at a time (Snijders, 2002) or by using a pure Metropolis algorithm (Hunter and Handcock, 2006) have also been suggested.

To calculate the MCMC-MLE, first consider the log-likelihood

$$\ell(\boldsymbol{\theta}) = \left\{ \sum_{T \subseteq C} \theta_T g_T(\mathbf{y}) - \kappa(\boldsymbol{\theta}) \right\}$$

where the normalizing constant has been written to emphasize its dependence on the unknown parameter vector, $\boldsymbol{\theta}$. The key idea behind this approach is that the parameter values that

maximize $\ell(\boldsymbol{\theta})$ are equivalent to those that maximize the log of the likelihood ratio

$$r(\boldsymbol{\theta}, \boldsymbol{\theta}^0) = \sum_{T \subseteq C} \left[(\theta_T - \theta_T^0) g_T(\mathbf{y}) \right] - \log[\kappa(\boldsymbol{\theta}) - \kappa(\boldsymbol{\theta}^0)] \tag{2.8}$$

for some fixed and constant $\boldsymbol{\theta}^0$. The difference of normalizing constants can be approximated by generating a sample of M graphs from the ERMG with parameter values $\boldsymbol{\theta}^0$ and noticing that

$$\exp[\kappa(\boldsymbol{\theta}) - \kappa(\boldsymbol{\theta}^0)] = E_{\boldsymbol{\theta}^0} \left\{ \exp \sum_{T \subseteq C} (\theta_T - \theta_T^0) g_T(\mathbf{y}) \right\}$$

Thus, an approximation of the likelihood ratio in (2.8) is

$$\hat{r}_M(\boldsymbol{\theta}, \boldsymbol{\theta}^0) = \sum_{T \subseteq C} \left[(\theta_T - \theta_T^0) g_T(\mathbf{y}_{\text{obs}}) \right] - \log \left\{ \frac{1}{M} \sum_{i=1}^{M} \exp \left[\sum_{T \subseteq C} (\theta_T - \theta_T^0) g_T(\mathbf{y}_i) \right] \right\}$$

and maximization of this equation gives and approximate to the MLE. Hunter and Handcock (2006) also proposed a method for approximating the standard errors of this estimate, performing a likelihood ratio test and extended this method to the curved exponential family which is sometimes necessary for the degeneracy-combating parameters discussed in the following section. Another extension of the MCMC-MLE approach of Geyer and Thompson (1992) is to estimate parameters while accounting for measurement error resulting from a partial observation of the network (Handcock and Gile, 2010).

The other main approach is based on the Robbins-Monro algorithm which solves equations of the form $E_{\boldsymbol{\theta}}[\mathbf{Z}] = 0$ for a random vector $\mathbf{Z}$ (Snijders, 2002). To estimate the parameters of an ERGM, the random vector takes the form $\mathbf{Z} = g(\mathbf{y}) - g(\mathbf{y}_{\text{obs}})$. The iteration step is

$$\hat{\boldsymbol{\theta}}^{(n+1)} = \hat{\boldsymbol{\theta}}^{(n)} - a_n D_n^{-1} Z(n)$$

where $\Pr(Z(n)|Z(1), \ldots, Z(n-1)) \equiv \Pr(Z|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(n)})$. The a_n is known as the gain sequence and is a sequence of positive values that converge to 0. If $a_n = 1/n$ and $\mathbf{Z}$ has an exponential family, the optimal choice of D_n is the derivative matrix (Kolaczyk, 2009). For ERGMs this is given by

$$D_{j,k} = \frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}$$

The three phase algorithm presented in Snijders (2002) is based on this approach and includes a check for validity and a method to compute the covariance matrix of the estimator.

Bayesian methods of estimation have not been extensively considered for ERGMs despite the fact that they are appropriate for quantifying uncertainty of the estimated model parameters and for formally comparing competing models (Caimo and Friel, 2011). Wong (1987) used an empirical Bayes approach to the p_1 model of Holland and Leinhardt (1981) which estimates the model parameters through a Newton-Raphson step and then uses an EM algorithm step to estimate the covariance parameters. The full algorithm cycles through both steps and hence was termed the EM-Newton algorithm. Gill and Swartz (2004) extended this to a fully Bayesian approach and also considered a model which includes a block effect parameter, but did not consider a model with a dependence structure beyond dyad-independence.

Computational complexity is the main reason that Bayesian estimation techniques have lagged behind for ERGMs. Usual MCMC algorithms are able to sample from posterior distributions as long as they are known up to a constant. An ERGM is doubly intractable because it is not possible to evaluate the normalizing constant of the posterior or the likelihood. This was overcome by adapting the exchange algorithm to network analysis (Caimo and Friel, 2011). This algorithm involves augmenting the data with $\mathbf{y}'$ and parameters with $\boldsymbol{\theta}'$ and sampling from the augmented posterior density

$$\pi(\boldsymbol{\theta}', \mathbf{y}', \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})h(\boldsymbol{\theta}'|\boldsymbol{\theta})\pi(\mathbf{y}'|\boldsymbol{\theta}')$$

where the posterior of interest, $\pi(\boldsymbol{\theta}|\mathbf{y})$ is a marginal distribution. The distribution $\pi(\mathbf{y}'|\boldsymbol{\theta}')$ is the same exponential family form as $\pi(\mathbf{y}|\boldsymbol{\theta})$ and the auxiliary density, $h(\boldsymbol{\theta}'|\boldsymbol{\theta})$, can be any arbitrary distribution with dependence on $\boldsymbol{\theta}$. The algorithm first samples the augmented data, $\mathbf{y}'$, and parameters, $\boldsymbol{\theta}'$, through a Gibbs sampler and then proposes a swap between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. The significance of this method is that in the Metropolis acceptance probability, all intractable normalising constants cancel. A further difficulty arises when trying to sample $\mathbf{y}'$ as it requires exact sampling (Hunter et al., 2012). Caimo and Friel (2011) propose sampling the augmented data using a "tie no tie" sampler which is computationally faster than a single dyad updating approach. In order to improve mixing and speed of convergence, the authors also

recommend using population MCMC which combines multiple, simultaneous chains. Another similar approach samples from an extended sample space using the linked importance sampler auxiliary (LISA) Metropolis-Hastings algorithm (Koskinen et al., 2010) . Like Handcock and Gile (2010), this work also focuses on networks that are only partially observed and thus the LISA algorithm is part of a larger sampling scheme that also samples from the distribution of missing values.

One area which has not been fully developed for ERGMs, or network analysis in general, is assessing the fit of the model to an observed network. Salter-Townshend et al. (2012) indicates four groupings of existing model assessment techniques: ground truth comparison, link prediction, model comparison, and graphical goodness-of-fit. The first two approaches are possible only when the network is assumed to be known and are often used to compare two generative methods. Model comparison includes likelihood based measures, such as AIC (Goodreau, 2007), or Bayesian model selection such as selecting the model that minimizes the expected predictive deviance (Gill and Swartz, 2004). Goodreau et al. (2009) warn against using the model comparison measures because they fail to indicate the manner in which the model is misspecified. Rather, the recommendation is for the graphical goodness-of-fit method (Hunter et al., 2008a). In this approach, a large number of graphs are simulated from the model with the estimated parameters. From these simulations a distribution of structural aspects are computed. Examples of structural aspects to consider are the number of realized edges, the number of triangles, or the average path length between pairs of nodes. The structural aspects can be the sufficient statistics included in the model or not. If the value of these structural aspects from the observed network appears to be a common value of the corresponding distribution obtained from the simulations, the the model is determined to be a good fit. The model is determined to fit poorly if the structural aspect was not a statistic included in the model. An observed value of a statistic included in the negpotential that is unlikely based on the distribution of simulated structural aspects is an indication of model degeneracy.

2.3.1.3 Degeneracy

Model or inferential degeneracy is a modeling issue that has been widely recognized in the application of ERGMs to empirical networks. The term degeneracy can be used to refer to various undesirable model behaviors. Most behaviors are a result of the model placing a large amount of probability on a very small subset of theoretically possible networks, which are often do not resemble the observed network and can include the complete (all edges realized) and/or empty (no edges realized) graphs. The phenomena has led to the definition of a "useful model" (Handcock, 2003a,b) as one for which the model places a large amount of probability on graphs that could reasonably be produced by the underlying process. Simulation, parameter estimation, and model assessment are listed as three interrelated capabilities a useful model will possess; features which are lacking or difficult to obtain in degenerate models (Handcock, 2003b) . Because an ERGM cannot be degenerate in the strict sense (Schweinberger, 2011), this behavior has also been referred to as near-degeneracy (Robins et al., 2007). Model degeneracy is not unique to ERGMs as a similar behavior has also been recognized in a more general class of models for interactive systems (Strauss, 1986) and is similar to long-range dependence observed in the Ising model (Snijders, 2002).

The inability of degenerate models to simulate reasonable graphs can lead to a failure to accurately estimate model parameters. As described in the previous section, parameter estimation for an ERGM requires the use of a sampling scheme due to the intractable normalizing constant. If a large amount of probability is placed on a few, disparate graphs, the algorithms can mix very slowly, moving very little for millions of steps (Schweinberger, 2011). A proposed solution to the slow mixing is the inclusion of a graph complement step in the Markov chain. At each step the value of all random variables is reversed with a small probability, e.g., for simple graphs $\mathbf{Y}^{(t+1)} = 1 - \mathbf{Y}^{(t)}$ (Snijders, 2002). Although this allows the algorithm to explore extreme modes of the distribution, the model is still degenerate. The issue of degeneracy is not a failure of the algorithm or statistical inference technique, but rather the underlying or stationary distribution, (Schweinberger, 2011; Snijders et al., 2006) or as a result of model misspecification (Goodreau et al., 2008). The MCMC-MLE algorithm will fail to find an estimate

when the subset of plausible graphs do not resemble the network of interest. The estimation technique requires that the model produce the observed values of the sufficient statistics (Snijders et al., 2006) and even with the graph complement step, the algorithm may continually jump over the observed values.

If an estimation algorithm does converge and estimates are obtained, simulated graphs from the fitted model could still fail to reproduce much of the graph structure observed in the realized network. One instance of this phenomena was explained via change statistics (Snijders et al., 2006). When an ERGM is simulated with a Gibbs sampler, the conditional distributions are stated as logistic regression with the change statistics playing the role of the independent variables, as shown in (2.7). The edge values are updated one at a time, either by cycling through all edges in the graph or through random selection. Consider a model with a moderately-sized, positive value of the triangle or any of the $k \geq 2$ star statistics. Changing one random variable to 1 could induce a large increase in the change statistics of other random variables, causing these edges to be realized with high probability, which in turn creates an increase in the change statistics for other edges, and so forth. This effect was termed an "avalanche of change" (Snijders et al., 2006) and would quickly force the algorithm to the complete graph with little probability of moving away. The graphical goodness-of-fit method (Hunter et al., 2008a), described in Section 2.3.1.2, was motivated by this commonly observed lack of fit. In addition, a proposed method to detect model degeneracy is to check if the observed model could possibly have been produced by the fitted model (Goodreau et al., 2009; Robins et al., 2007) .

Most of the work on diagnosing the cause of model degeneracy has identified a parameter space issue. Three of those approaches to identifying the offending parameter values will be discussed in detail below. Other potential causes include a large sample problem (Strauss, 1986; Strauss and Ikeda, 1990), where it can be proven that the expected density approaches 1 as the number of nodes increases to infinity. Likewise, an ERGM should not be applied to "large" graphs, with no specific definition of "large", as neighborhoods grow as a function of the number of nodes which has been identified as a culprit for degenerate models (Schweinberger and Handcock, 2012) . Another suggested reason for degeneracy is the dominating global

nature of an ERGM over local structure which is especially relevant when attempting to find the MPLE (Handcock, 2003a). Koskinen et al. (2010) proved that a model which contain nested, degenerate models will also be degenerate, which implies the issue of degeneracy can not be remedied by adding additional parameters into the model.

The first attempt to identify the subset of the parameter space which leads to degenerate models to be discussed relies the theory of discrete linear exponential families and the geometry of the parameter space. Handcock (2003a,b) extend results from Barndorff-Nielsen (1978, see Cor 9.6) and Rinaldo et al. (2009) makes use of Shannon's entropy to identify the problem areas. To explain the method in Handcock (2003a,b), let $g_{T1}(\mathbf{y}), \ldots, g_{T2}(\mathbf{y})$ represent the statistics chosen to be included in the negpotential function, $Q(\mathbf{Y})$ of (2.6), and let C represent the convex hull of the combination of possible values of the statistics computed from all possible graphs, $\{g_{\mathbf{T}}(\mathbf{y}) : \mathbf{y} \in \mathcal{G}\}$. Handcock (2003a) proved that the MLE will not exist if the observed values of the statistics, $g_{\mathbf{T}}(\mathbf{y}_{\text{observed}})$, fall on the relative boundary of C. This situation, he argues, occurs quite frequently in practice. Similarly, define C_1 to be the convex hull of the space formed by the statistic values computed from M simulated graphs, $\{g_{\mathbf{T}}(\mathbf{y}_1^*), \ldots, g_{\mathbf{T}}(\mathbf{y}_M^*)\}$. If $g_{\mathbf{T}}(\mathbf{y}_{\text{observed}})$ falls on the relative boundary of C_1, the MCMC-MLE will not exist. Therefore, if $C_1 \subset C$ there are observed statistics values for which the MLE exists, but the MCMC-MLE does not. This led to the proposed solutions of a Bayesian analysis where the prior distribution restricted all its mass to the non-offending areas of the parameter space.

Both Handcock (2003a) and Rinaldo et al. (2009) identified the problem region with a case study: a model with the density and 2-star parameter to a network with $n = 7$ nodes and a model with the density and triangle parameters to a network with $n = 9$ nodes, respectively. Handcock (2003a) identify the degenerate region through the convex hull method discussed above while Rinaldo et al. (2009) uses Shannon's entropy to quantify the amount of degeneracy where lower entropy corresponds to more degenerate parameter values. Both works concluded that the degenerate regions have a nicer, more identifiable form in the mean value parameterization rather than the natural parameterization of the exponential family. If we refer to the region of the parameter space for which degeneracy is not an issue as the effective parameter space (Handcock, 2003a), in both of the case studies this area was found to be much

smaller than the theoretical parameter space. Although the case studies represent only two possible ERGM specifications on unrealistically small networks, Rinaldo et al. (2009) claims that the results can be extended to any ERGM with nodes labeled arbitrarily and no node level information included.

Schweinberger (2011) also recognizes the issue of degeneracy as related to the discrete exponential family model, characterizing the issue as one of stability. He separately defines a stable distribution and a stable sufficient statistic. A distribution is stable if the maximum values of the negpotential function, $Q(\mathbf{y})$, over all possible graphs is bounded by some constant times the number of degrees of freedom, N, for large N. A stable sufficient statistic is one with a maximum value that is bounded in a similar manner. For ERGMs, the degrees of freedom, N, are equal to the number of possible edges, or $N = n(n-1)/2$ for a simple graph. The unstable distributions are characterized by excessive sensitivity and near degeneracy. Sensitivity is defined as distributions with unbounded nearest neighbor log-odds, where two possible graphs are considered nearest neighbors if the graphs are equivalent with the exception of a single edge. Six different ERGM specifications were analyzed to identify the regions of the parameter space which exhibited instability. All models contain a density term, the first three contain the 2-star, triangle, and both; the remaining three contain each one of the geometrically-weighted statistics, which will be discussed in detail below. Regions of instability are identified for all six models. The first two are stable only when they are generalized to a Bernoulli graph, i.e., when the 2-star and triangle parameters are zero. The area of stability for the third model with a density, 2-star, and triangle term is also very restrictive. For the models with the geometrically weighted terms, the area of stability is non-negligible, providing some optimism; however, care is still needed to avoid the unstable parameter regions.

Identification of degenerate parameter space regions of an ERGM has also been examined by researchers outside the disciplines of statistics and sociology. Bhamidi et al. (2008) examine the mixing time for an MCMC of an ERGM specified with a density and triangle parameter. A high and low temperature regime are defined and identified for the parameter space. The authors showed that the algorithm mixes exponentially slow in the low temperature regime and as $\Theta(n^2 \log n)$ in the high temperature regime. However, the authors also showed that

models with parameter values in the high temperature regime are asymptotically independent and thus are not appreciably different from the Erdös-Rényi graph. This coincides with the region of instability found for the model by Schweinberger (2011) for the same ERGM specification. Park and Newman (2004, 2005) investigate the degeneracy issue with analysis techniques from statistical physics for two ERGM specifications: the specification with a density and two-star parameter, called the two-star model, and the specification with a density and transitivity parameter, called the clustering model. For the two-star model, the authors used the Hubbard-Stratonovich transform and saddle-point expansions, to determine the region of the two-parameter space where degeneracy occurs. High and low density phases which are separated by a coexistence region are identified in a phase diagram. This coexistence region corresponds to a symmetry-broken phase. The symmetry breaking that describes the coexistence phase results in similar behavior as stability (Schweinberger, 2011). The separation of these three phases corresponds to a conventional continuous phase transition. As a result of this analysis the authors concluded that the degeneracy problem, at least in this particular ERGM, is analogous to a phenomenon in physics known as phase separation. The coexistence region was also identified for the clustering model (Park and Newman, 2005). For this particular model, the degenerate region corresponds to parameter values that indicate a moderate number of triangles. The finding led the authors to conclude that without some augmentation, the clustering model will never adequately describe a real-world network.

In addition to the case studies mentioned above, special attention has been given to degenerate ERGM specifications that include a term for transitivity, and to a lesser extent to those with k-star terms. Inclusion of a term that accounts for transitivity has been shown to be important because the closure of triangles is a main feature that separates realized networks, at least for social networks, from those generated independently and at random (Snijders et al., 2006). However, incorporation of a term that successfully reflects the amount of transitivity in the observed network has been a difficult task. This problem was explored through simulations which demonstrated for an ERGM with only a density and triangle term, an increase in the value of the triangle parameter does not correspond to a smooth increase in the number of triangles in the simulated graph (Robins et al., 2007). Instead what occurs is a tendency to

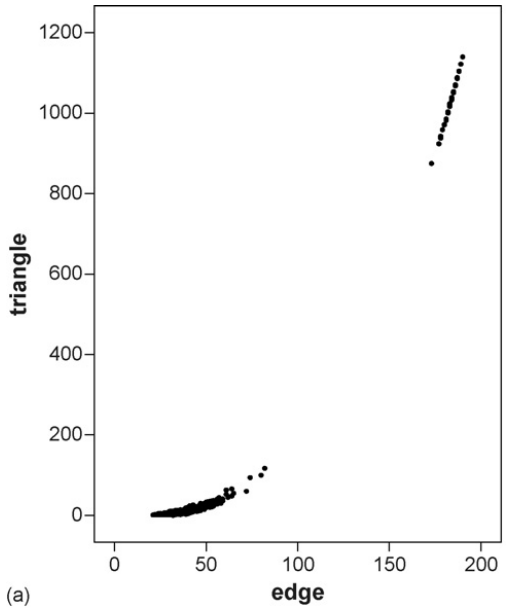


Figure 2.3 Scatterplot of number of edges against the number of triangles from a simulation study conducted by Robins et al. (2007) for an ERGM with density parameter fixed at -1.5 and triangle parameter ranging from 0 to 1.

resemble either a high or low density Erdös-Rényi graph, in agreement with the conclusion of asymptotic independence of the same model by Bhamidi et al. (2008). The number of triangles in the simulations occur uniformly throughout the graph and form as a function of less or more edges with a dramatic jump (Robins et al., 2007), shown in Figure 2.3, between the two extremes. Therefore, the model is unable to adequately describe a graph with a moderate number of triangles. If the parameter estimate is obtained with MCMC-MLE, the parameter will be estimated to be the value at which the jump occurs. The probability distribution of the statistic will be bimodal with a combination of the low density graphs to the left of the jump and the high density graphs to the right (Snijders et al., 2006).

One proposed reason for the difficulty in the representing transitivity with only a triangle parameter is the existence of other effects contributing to the formation of triangles (Snijders et al., 2006). For example, three relationships could form independently between three students in the same class, not because of a shared relation, but because of the shared covariate value.

Even if a term is included in the model to explain the covariate effect, the two processes are not acting on the edges independently although they are modeled as such (Goodreau et al., 2009). Another reason for the difficulty with modeling transitivity through a count of triangles is the manner in which the model introduces triangles does not correspond to how they appear in observed networks. The model wants to place the increasing number of triangles uniformly through the network, where it has been observed that groups of triangles tend to form dense "clumps" of triangles (Robins et al., 2007) . These clumps of edges are not completely connected (Snijders et al., 2006) and thus the more triangles to which an edge is a part of, the less likely it is to be part of a new triangle. These observations have led to the formation of a new specification for representing transitivity, the alternating k-triangle statistic.

The alternating k-triangle statistic (Snijders et al., 2006) was proposed to model transitivity but avoid the avalanche effect that leads to degeneracy. The motivation for this statistic is transitivity in observed networks is important but complex, and that a statistic that merely counts the number of triangles is overly simplistic. A k-triangle is a set of k triangles that share a common edge, often referred to as the base; see Figure 2.4 for an example 5-triangle. The formula for the alternating k-triangle statistic is

$$AKT_\lambda(\mathbf{y}) = 3T_1 + \sum_{k=2}^{m-2} (-1)^{k+1} \frac{T_k(\mathbf{y})}{\lambda^{k-1}}$$

where T_k is the number of k-triangles. The increasing denominator gives decreasing probability to higher-order triangles as per the empirical observation noted above. The value of λ controls the type of transitivity. Larger values of λ lead to smaller probability given to higher-order triangles where smaller values lead to a localized effect (Hunter and Handcock, 2006). The alternating signs of increasing terms also aim to prevent large cliques of edges.

The alternating k-triangle statistic is included in an ERGM as a term in the negpotential function with a single parameter coefficient. The value of λ can either be considered known or estimated. If this value is to be estimated, the ERGM is no longer within the standard exponential family as the negpotential is not a linear function of the parameters. Hunter and Handcock (2006) present the details for estimation in a curved exponential family for an ERGM where λ is to be estimated using MCMC approximation to the likelihood. This work

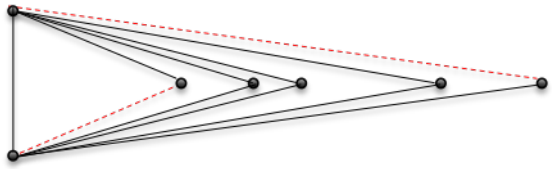


Figure 2.4 An example 5-triangle

does recommend estimating this parameter unless theoretical considerations imply a particular value.

Snijders et al. (2006) show that an ERGM with the alternating k-triangle statistic satisfies the partial conditional dependence of Pattison and Robins (2002). Thus, the dependence graph for this particular model is realization dependent. The authors argue for a more generalized dependence structure than the Markovian dependence claiming that edges that are not incident could possibly be conditionally dependent. Robins et al. (2007) claim that considering cliques of size greater than three is necessary to avoid degeneracy. The alternating k-triangle statistic allows for edges that, if realized, would create a four-cycle to be conditionally dependent. To make this connection more clear, the two red, dashed lines in Figure 2.4 are not incident, but would be conditionally dependent under the partial conditional dependence structure.

Inclusion of the alternating k-triangle statistic makes interpretation of model parameters more difficult (Snijders et al., 2006), and it does not lead to a simple representation of dependency (Schweinberger and Handcock, 2012). As a demonstration of the parameter interpretation, consider the case study of collaboration relationships between $n = 36$ partners in a law firm (Snijders et al., 2006). One model fit to this data included the alternating k-triangle statistic and multiple attribute-based statistics. A significant coefficient for the k-triangle term is interpreted as evidence of a triangle formation process beyond what could be explained by considering only attributes. When an additional term is included to to represent the degree distribution, the significant k-triangle parameter indicates that transitivity is not the result of popularity of nodes. A significant and positive k-triangle parameter can be interpreted as indication of a core-periphery structure resulting from transitivity, rather than popularity

(Robins et al., 2007). Although the k-triangle statistics appears "contrived", it is argued that a contrived statistic is necessary due to the complex processes that work together to create the static view of the network.

An alternative formulation for the k-triangle statistic is based on shared partner statistics which can lead to a more clear interpretation of parameters (Hunter, 2007). The edgewise shared partner statistic, $EP_k(\mathbf{y})$ counts the number of edges that are realized with both nodes connecting to exactly k other nodes. The alternating k-triangle statistic is then equivalent to the geometrically-weighted edgewise shared partner statistic,

$$\text{GWESP}_\theta(\mathbf{y}) = e^\theta \sum_{i=1}^{m-2} \left\{1 - (1 - e^{-\theta})^i\right\} EP_i(\mathbf{y})$$

where $\theta = \log \lambda$ in the original formulation. This parameterization is particularly useful when the value of λ is to be estimated because of the desired restriction to positive values. With this formulation of the statistic, a significant, positive parameter would be interpreted as the more edges two nodes have in common, the less the motivation is to form more common edges.

The alternating k-triangle statistic was novel in its approach because rather than counting local configurations, the statistic attempts to summarize an entire distribution of subgraph counts (Hunter et al., 2008b). Two other distribution-summarizing statistics have with the same goal of decreasing the effects of degeneracy (Snijders et al., 2006). The first is the alternating k-star statistic

$$AKS_\lambda(\mathbf{y}) = \sum_{k=2}^{m-1} (-1)^{-k} \frac{S_k(\mathbf{y})}{\lambda^{k-2}} \tag{2.9}$$

where $S_K(\mathbf{y})$ counts the number of k-stars. The alternating k-star statistic is an attempt to model the degree distribution. The Markov model of Frank and Strauss (1986) proposed including terms for all k-stars, restricting it to only the first two to decrease the number of parameters from $n - 1$ to 2. The alternating k-star statistic also restricts the number of parameters from $n - 2$ to 2, if the value of λ is to be estimated. Therefore, inclusion of the $ASK_\lambda(\mathbf{y})$ can be seen as modeling all k stars but maintaining a reduction of the parameter space. As the issue of degeneracy has been identified as a parameter space issue, limiting attention to a subset of the parameter space can lead to an increased probability that the MLE will exist.

The other statistic is the alternating k-twopath statistic

$$AKP_\lambda(\mathbf{y}) = \sum_{k=1}^{m-2} (-1)^{-k} \frac{P_k(\mathbf{y})}{\lambda^{k-2}}$$

where $P_k(\mathbf{y})$ counts the number of k-twopaths which are similar to a k-triangle, but without the base. The purpose of the alternating k-twopath statistic is to be used in conjunction with the alternating k-triangle statistic. If both statistics are included the k-triangle parameter can be interpreted exclusively as transitivity, rather than as transitivity and the necessary conditions for transitive closure. Equivalent geometrically weighed versions of $ASK_\lambda(\mathbf{y})$ and $AKP_\lambda(\mathbf{y})$ also exist (Hunter, 2007) .

A more recent adaptation to counter degeneracy is a hierarchical ERGM (Schweinberger and Handcock, 2012). In this work, the issue of model degeneracy is attributed to large and growing neighborhoods, where a neighborhood of an edge refers to the set of edges on which the original edge is conditionally dependent. For a simple graph with Markovian dependence where all $\binom{n}{2}$ edges are possible, each potential edge has a neighborhood of size $2(n-2)$. Thus, as the number of nodes increases, so does the size of the neighborhoods.

The first feature of a hierarchical ERGM is a partitioning of the nodes into K local "neighborhoods". These "neighborhoods" are non-overlapping, although edges can form between them. This interpretation of "neighborhood" is more consistent with the concept of a community; thus, in order to avoid confusion, the rest of this description will reflect this terminology. The community structure is not necessarily an observable feature of the graph, nor are the number of communities required to be known a priori. The second feature of the proposed model is local dependence and global independence. The distribution function of the graph is separated into a within- and between-community probability mass function (PMF). If $\mathbf{Y}_{(kl)} = \{Y_{ij} : i \in \mathcal{N}_k, j \in \mathcal{N}_l\}$ represents the edges that could form between nodes in community k and nodes in community l, the conditional probability function for the entire graph $\mathbf{Y}$ given the community membership, $\mathbf{X}$, is written as

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) = \prod_{k=1}^{K} P(\mathbf{Y}_{(kk)} = \mathbf{y}_{(kk)}|\mathbf{X} = \mathbf{x}) \prod_{k<l}^{K} P(\mathbf{Y}_{(kl)} = \mathbf{y}_{(kl)}|\mathbf{X} = \mathbf{x})$$

Each PMF can include a different set of sufficient statistics and thus parameters to estimate and each set of parameters for the between- and within-community PMF will induce a different

dependence structure on a subset of the graph. A Bayesian inference method is presented to estimate the number of communities, node community membership, and all sets of parameters. This computationally expensive procedure requires an approximation to the prior and an exchange algorithm to sample from the posterior.

2.3.2 Latent Variable Models

The subclass of probabilistic network models known as LVMs encompasses a broad class of models that are hierarchical in nature. The common thread between LVMs is the edge variables are specified as conditional distributions and considered conditionally independent given some latent variable, such as block membership or position within a social space. Any model which treats the nodes as exchangeable can be represented as a latent variable model (Salter-Townshend et al., 2012). In contrast to an ERGM where the interpretation of the model identifies highly probably edges, the focus of an analysis from a latent variable model is on the role and position of individual nodes of the graph. The goal of incorporating a latent structure is to account for some of the variability of the topological features in the observed network (Goldenberg et al., 2010). While an ERGM is really a single class of models with proposed modifications, the latent variable category encompasses a variety of models. Those that will be discussed are the random effects models, latent block models, and latent space models.

2.3.2.1 Random effects models

The simplest random effects model is the p_2 model (Van Duijn et al., 2004), a random effects version of the p_1 model of Holland and Leinhardt (1981). Both the p_1 and p_2 models were developed for directed social networks and contain parameters to describe density, reciprocity, and node-level expansive/productive effects and popular/attractive effects. In the p_1 model the node-level parameters are modeled as fixed effects with a potentially unique value for each node. For identifiability purposes, constraint must be placed on this set of parameters. In contrast, the p_2 model treats node-level parameters as crossed random effects and aims to estimate the parameters of the underlying distribution from which they could have reasonably been drawn (Goldenberg et al., 2010). Covariates can be incorporated into the p_2 model, at the

node-level with additional modeling of the attractiveness and productivity parameters and/or by modeling the density and reciprocity parameters as functions of dyad-level covariates. The coefficients of the covariates are modeled as fixed effects and hence the p_2 model can be viewed as a generalized linear mixed model (Zijlstra et al., 2009). The remaining node-level variability not explained by the covariates is incorporated into the correlated random effects. If there is no covariate information, the p_2 model is merely a more parsimonious version of the p_1 model (Salter-Townshend et al., 2012).

Parameter estimation of the p_2 model was originally conducted via an Iterative Generalized Least Squares algorithm (Van Duijn et al., 2004). Because p_2 is a nonlinear model, the algorithm requires a linearization step based on a Taylor series expansion of the likelihood function around the current estimate of the parameters. Zijlstra et al. (2009) introduced three Bayesian methods for parameter estimation of the p_2 model. All approaches implement a Gibbs sampler, with separate updating steps for the random effects, covariance matrix, and fixed parameters. The full conditional distributions of the random and fixed effects cannot be directly sampled; thus, the three separate approaches explore three different proposal distribution for the required Metropolis-Hasting steps.

The other area in which random effects have been utilized for network analysis is by incorporating them into a generalized linear model (GLM) framework (Hoff, 2003). The purpose of the random effects are to model higher-order dependence. For example, within-node dependence can be represented as random intercepts on the random effects terms. Traditionally, a GLM assumes observations are conditionally independent given regression coefficients, or fixed effects, whereas the model presented in Hoff (2003) assumes conditional independence given the random effects terms. An estimation scheme is also presented where the regression coefficients, fixed effects, and covariances of the random effects are estimated with a standard Bayesian analysis.

The generalized bilinear regression model (Hoff, 2005) is a direct extension to the generalized linear mixed-effects model discussed above. The extension is to include a bilinear effect into the error structure to incorporate third-order dependence, or dependence between triples of random variables. This bilinear effect is also referred to as a reduced-rank interaction term and is an

inner product of latent characteristic vectors. Latent characteristic vectors are then modeled as independent K-dimensional multivariate normal distributions with mean zero and diagonal covariance matrices. This inner product can be viewed as a mean zero random effect. As an application of this method, Hoff (2005) analyze a valued network of international relations in central Asia.

The methods of Hoff (2005) are applied in Hoff and Ward (2005) to a network of bilateral trade in an analysis to determine how it is affected by various country attributes, such as capitalism, conflict, cooperation and democracy, among others. Although not statistically novel, through this approach the authors were able to explain three-fourths of the variability in the network by accounting for second- and third-order dependence over the standard gravity model that had traditionally been able to explain only one-half. In addition, through the random effects, correlations were modeled between importer, exporter, and dyadic relations, where they had previously been modeled as independent. One of the more interesting result of the analysis was that bilateral trade was not significantly impacted by conflict between two countries, although cooperation between two countries led to a significantly increased amount of bilateral trade.

2.3.2.2 Latent blockmodels

In the most general sense, blockmodels are models which classify the nodes of the graph into groupings, often referred to as blocks. There are two, potentially overlapping, concepts of block structure. This first is that edges are more likely to form between nodes classified within the same block than between nodes in disparate blocks. This is the approach adopted by the computer science community where a block is often referred to as community. One of the main open challenges in this field is that of "community detection", or uncovering these groups of nodes (Goldenberg et al., 2010). The second definition of block is motivated by the notion of structural equivalence (Fienberg, 2012). Two nodes are defined to be structurally equivalent if the relations between those nodes and all other nodes in the graph is equivalent. A relaxation of the this concept often referred to as stochastic equivalence can be descried as two nodes that relate to similar nodes in a similar manner. With this second definition, the purpose of

blocking is to capture the main structural features of the network (Snijders and Nowicki, 1997). Within the statistics literature models that incorporate block structure have been referred to as stochastic blockmodels. This work can be partitioned into either a priori or a posteriori.

The a priori blockmodels were introduced by Wang and Wong (1987) and are presented as an extension of the p_1 model of Holland and Leinhardt (1981). These models do not fit under the heading of "Latent" Blockmodels as one assumption is that the block structure is observed. The block structure is incorporated into the p_1 model by inclusion of a block-specific parameter that accounts for block membership and adjusts the other parameters in the model, such as expansiveness and popularity, for this membership.

The obvious disadvantage of the stochastic blockmodels of Wang and Wong (1987) is the requirement to know the block membership a priori. When this information is unknown, blockmodels of the second type, a posteriori, are used to infer the group membership. An early attempt was made by Wasserman and Anderson (1987) who fit the p_1 model to data, grouped nodes based on the estimates of the productivity and popularity parameters, and then, given the group structure, fit a pair-dependent stochastic blockmodel. The pair-dependent stochastic blockmodel specifies the joint distribution of the edge random variables given parameter values and the block membership.

Extensions to the pair-dependent approach were developed for undirected, binary graphs with only two blocks (Snijders and Nowicki, 1997) and directed, weighted networks and an arbitrary number of blocks (Nowicki and Snijders, 2001). Both approaches include two assumptions: the number of blocks is known and given block membership, edge probabilities are independent. Block membership is inferred from the pattern of edges and estimated through a Gibbs sampler which alternatively samples the parameters of the model and the block membership of the nodes. Membership probabilities and parameters are assessed through posterior distributions. Parameter identifiability is an issue common for mixture models as the model can only distinguish between different partitions, not the distinct labeling of them. This leads to distinct parameter values resulting in the same probability distribution. One proposed solution is to impose order restrictions on the block probabilities; however, the method was shown to lead to poor group identification when probabilities of different blocks are similar. There-

fore, the approach taken in Nowicki and Snijders (2001) is to restrict attention to posterior distributions of functions which are invariant to relabeling.

Specification of blockmodels requires two components: the block model itself and an index of the nodes and the blocks to which they belong (Goldenberg et al., 2010). An extension of the second component is the mixed-membership stochastic blockmodel. For the previously mentioned stochastic blockmodel, there is a one-to-one mapping from node to block. In contrast, the mixed-membership model specifies an array of memberships for each node, so that the block to which the node belongs depends upon the node with which it would potentially join, thus, group membership is "context dependent" (Fienberg, 2012). For directed graphs, each node's membership array is of length $2n - 2$. The motivation for this model is that the parameters of the block model describe the global features of the graph, while the membership arrays capture the node-specific patterns (Goldenberg et al., 2010).

2.3.2.3 Latent space models

The main idea of the latent space models is that the nodes of the graph can be represented in a low-dimensional, latent space and the distance between two nodes affects the probability that an edge will form between them. Given the position of the nodes, or rather the distance between them, the probability of the edges are conditionally independent. One purpose of this latent space is to incorporate the effects due to unmeasured covariates (Vivar and Banks, 2012).

The latent space model is presented in Hoff et al. (2002). In this work the authors proposed two models: the distance model and the projection model. The models differ in the way in which the latent space is incorporated into the probability. Euclidean distance is used by the distance model, although the authors point out that any metric could be used. A metric over the latent space inherently incorporates reciprocity, through the symmetry requirement of a metric, and transitivity, through the triangle inequality. Due to its ease of interpretability, the distance model has been more widely used in practice (Salter-Townshend et al., 2012). The projection model incorporates the latent positions of two nodes through the projection of the position of one node onto the direction of the other. This is a measure of similarity two nodes share with respect to some characteristics and depends on the angle the positions create in

Bilinear latent space (Salter-Townshend et al., 2012). The projection model has been shown to be most appropriate when the graph is strongly asymmetric. If additional covariate information is available, the model also allows this information to be explicitly incorporated.

As originally conceived, the latent space model is able to capture three important features of networks: transitivity, reciprocity, and homophily of attributes. One aspect of networks that the model cannot account for is that of clustering, also referred to as community structure. To incorporate this feature, the Latent Position Cluster Model was developed (Handcock et al., 2007). This model explicitly incorporates clusters in the latent space using a mixture of spherical Gaussian distributions on the positions in the latent space. Krivitsky et al. (2009) further extended this model by combining the approach of Handcock et al. (2007) and Hoff (2005) into the Latent Cluster Random Effects Model. This model is able to account for heterogeneity of the nodes, in addition to the four previously mentioned features of networks. Node-specific random effects are added to represent sociality effects in an undirected network, and for directed networks a sender and receiver effect is specified for each node. The dimension of the latent space and the number of clusters is not assumed to be known a priori. Choosing these values is viewed as a model selection problem.

All of the approaches mentioned above can be cast as a generalized linear model as defined by three features: the error model, $\Pr(Y_{ij})$, the linear model, η_{ij}, and the link function $g(\mu_{ij}) = \eta_{ij}$ where $\mu_{ij} = E(y_{ij})$. The models described above are presented as an analysis method for unweighted graphs and thus the error model is $\Pr(Y_{ij}) = \text{Bernoulli}(\mu_{ij})$ with link function $g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right)$. It is the linear model, η_{ij}, to which the additional terms are added. As an example, if Z_i represents the latent position of node i and $x_{i,j}$ a dyad-level covariate between nodes i and j, the link function, or conditional log-odds given model parameters $(\alpha, \boldsymbol{\beta})$ and latent positions, for the projection model (Hoff et al., 2002) can be written as

$$\eta_{ij} = \alpha + \boldsymbol{\beta}' x_{i,j} + \frac{Z_i' Z_j}{|Z_j|}$$

The extension to consideration of clustering is to assume the Z_is are realizations from a finite mixture of multivariate distributions and the extension to account for node heterogeneity is accomplished by adding random effects terms to the linear model. When the latent space

model is constructed in this manner the extension to a weighted graph is natural by specifying a different error model and link function. This approach was demonstrated on a network where the edges are counts and thus the error model used is $\Pr(Y_{ij}) = \text{Poisson}(\mu_{ij})$ with link function $g(\mu_{ij}) = \log(\mu_{ij})$ (Krivitsky et al., 2009).

Estimation of the unknown values of the latent space models can be accomplished through either maximum likelihood or a Bayesian analysis of posterior distributions (Handcock et al., 2007). The maximum likelihood is conducted in two stages. The first stage estimates the distances between the nodes in the latent space, a relatively simple task as the log-likelihood is a convex function of the distances. Multidimensional scaling is then used to determine the positions of the nodes. The second stage involves determining the MLE of the model parameters, $(\alpha, \boldsymbol{\beta})$, and the parameters of the Gaussian mixture model, when necessary. Although this approach is fast and simple, the two separate stages imply that information about the first stage is not used in the estimation of the parameters in the second, and vice versa. A more common approach is to estimate all unknowns simultaneously with MCMC sampling.

The Bayesian approach to estimation utilizes a Gibbs sampler to cycle through the model parameters, latent positions, and group memberships, if clustering is of interest. For most parameters, it is possible to specify a conjugate prior and sample directly from the posterior distribution, but not all, and thus some Metropolis-Hastings steps are also necessary. Although this approach incorporates all information into each step, it is more computationally intensive than its frequentist counterpart, and does not scale well for large graphs. To update the latent position of each of the n nodes requires the calculation of $n - 1$ terms of the log-likelihood and the updating of the model parameters $(\alpha, \boldsymbol{\beta})$ requires all $O(n^2)$ terms be computed (either $n(n-1)$ for directed or $0.5 * n(n-1)$ for undirected) (Raftery et al., 2012). In practice, this estimation technique has been infeasible for $n > 1000$.

Due to the inability of the Bayesian estimation technique to adequately scale to large graphs, alternative methods have been proposed. First, an approximation to the log-likelihood is motivated by the approach of case-controls studies (Raftery et al., 2012). It makes use of the fact that networks are generally very sparse, i.e., a row or column in the adjacency matrix will contain or order of magnitude more 0's than 1's. In the analogy, the 1's are the cases

Table 2.2 Table of notation used within the literature review of Chapter 2.

V	Set of nodes/vertices
E	Set of edges
n	Number of nodes/vertices
m	Number of edges
$\mathbf{Y}$	$n \times n$ adjacency matrix
Y_{ij}	Random variable representing possible edge ij
$d(i)$	Degree of node i
γ	Power Law exponent
C	Clustering coefficient
$\mathcal{G}$	Set of all possible graphs
G	a graph
$\boldsymbol{\delta}_g(\mathbf{y})_{ij}$	Change statistic

and the 0's are the controls and the proportion of non-ties, i.e., 0's, are approximated through sampling. This approach reduces computation time from $O(n^2)$ to $O(n)$. Variational methods have also been used to approximate the Bayesian estimation of latent space models (Hunter et al., 2012).

Another issue with a Bayesian analysis of the latent space models is that the likelihood is invariant to rotation and reflection of the nodal positions and if a Euclidean space is used, also to translation (Hunter et al., 2012). In addition, if clustering is considered, the likelihood is invariant to the relabeling of the clusters, or the "label-switching problem" as it is known in mixture models (Handcock et al., 2007). The proposed solution to these issues requires a post processing of the MCMC output. The former is addressed by performing a Procrustean transformation on the posterior draws so that the result is close to a reference configuration, typically the MLE of the positions centered at the origin (Hoff et al., 2002). Alternatively, the framework proposed by Handcock et al. (2007) aims to correct both identifiability issues by minimizing the Bayes risk relative to a Kullback-Leibler divergence. In this approach the goal is to find a configuration that gives edge values closest to the posterior predictive distribution.

CHAPTER 3. A LOCAL STRUCTURE MODEL FOR NETWORK ANALYSIS

A paper submitted to *Statistics and Its Interface*

Emily M. Casleton, Daniel J. Nordman, Mark S. Kaiser

Abstract

The statistical analysis of networks is a popular research topic with ever widening applications. One common statistical modeling approach for this purpose are exponential random graph models (ERGM), which specifies a model through interpretable, global network features. In this paper we introduce a new class of models for network analysis, local structure graph models (LSGM). In contrast to an ERGM, a LSGM specifies a network model through local features and allows for an interpretable and controllable local dependence structure. In particular, LSGMs are formulated by a set of full conditional distributions for each network edge, e.g., the probability of edge presence/absence, depending on neighborhoods of other edges. Additional model features are introduced to aid in specification and to help alleviate a common issue (occurring also with ERGMs) of model degeneracy. The proposed models are demonstrated on a network of tornadoes in Arkansas where a LSGM is shown to perform significantly better than a model without local dependence.

3.1 Introduction

Applications of networks appear in a wide variety of disciplines. For example, sociologists use graph models to represent social networks, economists have used networks for studying relations between countries (Hoff and Ward, 2005), biologists to represent brain connectivity

(Sporns et al., 2004; Simpson et al., 2011), zoologists for examining animal social behavior (Lusseau, 2003), and computer scientists for representing connections on the internet.

Much literature is devoted to algorithmic construction methods with a goal to quickly and accurately simulate networks mimicking certain properties of interest (Leskovec et al., 2010). Such methods are often not statistical in nature in the sense that the algorithms involve no probability models producing tractable likelihood inference. In contrast, some network analysis approaches allow for explicit probabilistic modeling and related likelihood inference. In a review, Hunter et al. (2012) categorize probabilistic modeling of networks into the exponential random graph models (ERGMs) and latent variable models (LVMs). Of these two, ERGMs have been more widely used and extensively studied. Their popularity can be attributed to the ability to incorporate graph topology as terms of a joint (log-linear) distribution that allows for complex dependencies (Kolaczyk, 2009). Although ERGMs allow for complex dependencies, such dependencies are typically induced rather than directly specified. That is, dependencies in ERGMs are a consequence of graph topologies chosen to be included in the joint distribution. Latent variable models encompass a broad class of models that are hierarchical in nature. Here variables representing edges are commonly specified as having conditional distributions that are conditionally independent given some latent variable defined on the nodes, such as group membership (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001) or position within a social space (Hoff et al., 2002; Handcock et al., 2007).

In this chapter, we introduce an approach to specifying a model for network analysis we call local structure graph models (LSGMs). As a key characteristic, the LSGMs begin model formulation based on the set of full conditional distributions for each potential edge in the network, the distribution for the presence/absence of an edge given the outcomes of all other potential edges. As a further critical characteristic, each conditional distribution is specified in terms of a flexible neighborhood structure, explicitly identifying a set of other network edges on which an edge of interest is "locally" dependent. Under certain conditions, conditional specifications and neighborhood definitions allow construction of a global or joint probability model for the network, having a dependence structure which is interpretable and introduced in a controlled, local manner.

As a consequence of its formulation, LSGMs have characteristics of *both* ERGMs and LVMs, a feature of what has been called "the next generation" of network analysis (Snijders, 2007). Similarly to LVMs, LSGMs are specified through conditional distributions. However, the conditional distributions in the LSGM are not defined in terms of latent variables for nodes, such as group memberships, but rather in terms of neighborhoods involving other network *edges*. Hence, potential network edges are conditionally dependent on other edges belonging to a neighborhood. Like ERGMs, LSGMs result in joint distributions that have a Gibbsian form for random graphs. But in LSGMs the joint distribution results from a set of specified conditioned distributions for edges, while in ERGMs the joint is formulated directly. Consequently, the dependence structure of an ERGM is often induced, while that of the LSGM can be more directly and explicitly defined.

Because the joint distributions of LSGMs are similar to those for ERGMs, some details of ERGMs are discussed in Section 3.2. The main features of LSGMs, involving conditional specifications and neighborhood definitions, are detailed in Section 3.3, along with a numerical demonstration of a model. Two additional features of LSGMs, the ability to simply incorporate potential spatial information about nodes, and the definition of a "saturated graph," are also introduced in Section 3.3. These features can help keep the potential sizes of LSGM neighborhoods manageable which is useful for minimizing model degeneracy issues. In Section 3.4, a LSGM is applied to an example network consisting of tornado outbreaks in the state of Arkansas. Two simulation-based model comparison techniques are also presented in this section and used to compare the fit of the LSGM to that of a model lacking local dependence. It is shown that the Arkansas tornado network does exhibit significant local dependence which is appropriately described by the LSGM. Section 4.6 provides some concluding remarks.

3.2 Exponential Random Graph Model (ERGM)

A network, or graph, is defined by a set of n nodes and m edges, where the networks of interest here are undirected and simple, with unweighted edges and no self-loops. To construct a random graph model, assign to each of the $\binom{n}{2}$ possible edges a binary random variable $Y(\mathbf{s}_i)$, where the marker $\mathbf{s}_i = \{c_i, r_i\}$ indicates the two nodes, denoted as c_i and r_i, that a potential

edge would join. Edge values are collected into $\mathbf{Y}$, an $n \times n$ adjacency matrix, and each entry designates the presence, $y(\mathbf{s}_i) = 1$, or absence, $y(\mathbf{s}_i) = 0$, of an edge between each node pair in the graph. For undirected, simple networks, $\mathbf{Y}$ will be symmetric with $\text{diag}(\mathbf{Y}) = 0$. A realization of the network will be represented as $\mathbf{y}$.

Specification of an ERGM involves identifying the number of elements of $\mathbf{Y}$ that correspond to edges of certain types, which are often called topological features of the graph. For example, the well-studied triad model of Frank and Strauss (1986) includes the topological features of density, or the expected proportion of realized edges, 2-stars, and triangles. Let the classes of edge types to be included in an ERGM be indexed by $j = 1, \ldots, q$. For any possible realization $\mathbf{y}$, let $g_j(\mathbf{y})$ denote the number of occurrences of edge class j present in $\mathbf{y}$. The joint distribution of $\mathbf{y}$ is then specified as

$$\Pr(\mathbf{Y} = \mathbf{y}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{ \sum_{j=1}^{q} \theta_j g_j(\mathbf{y}) \right\} \tag{3.1}$$

where θ_j is a model parameter corresponding to topological graph feature of type $j = 1, \ldots, q$.

The summation

$$Q(\mathbf{y}) = \sum_{j=1}^{q} \theta_j g_j(\mathbf{y}) \tag{3.2}$$

is often referred to as the negpotential function, or Hamiltonian, and $Z(\boldsymbol{\theta}) = \sum_{\mathbf{y}} \exp\{Q(\mathbf{y})\}$ is a normalizing constant for the discrete distribution in (3.1).

Hence, specification of an ERGM involves identifying topological graph features of interest for defining the negpotential (3.2) with statistics, $g_j(\mathbf{y})$, as counts of such features. Let N be the set of all edges which share a common node and C be the set of all edges which could potentially form a triangle. The negpotential function for the triad model of Frank and Strauss (1986) can be written as

$$Q(\mathbf{y}) = \rho \sum_{i} y(\mathbf{s}_i) + \sigma \sum_{\mathbf{s}_i, \mathbf{s}_j \in N} y(\mathbf{s}_i) y(\mathbf{s}_j) + \tau \sum_{\mathbf{s}_i, \mathbf{s}_j \mathbf{s}_k \in C} y(\mathbf{s}_i) y(\mathbf{s}_j) y(\mathbf{s}_k) \tag{3.3}$$

where the sufficient statistics correspond to topological features as the number $g_1(\mathbf{y}) = \sum_i y(\mathbf{s}_i)$ of edges, the number $g_2(\mathbf{y}) = \sum_{i,j} y(\mathbf{s}_i) y(\mathbf{s}_j)$ of 2-stars, and number $g_3(\mathbf{y}) = \sum_{i,j,k} y(\mathbf{s}_i) y(\mathbf{s}_j) y(\mathbf{s}_k)$ of triangles. In (3.3), ρ represents a density parameter for the graph, σ represents a clustering parameter (Frank and Strauss, 1986), and τ is a parameter for transitivity. Hence, the

dependence structure of an ERGM is defined by the choice of graph features included in the specification (3.1) or (3.2); see Goodreau (2007) for details on choosing topological configurations.

Initially, ERGMs included parameters to represent the density, transitivity, and k-stars of the network (Frank and Strauss, 1986), of which the triad model (3.3) is a special case. This set of parameters leads to a Markovian dependence structure where two potential edges are conditionally dependent if they share a common node. ERGMs can also be expanded to incorporate more complicated graph topologies (Wasserman and Pattison, 1996), exogenous covariate information (Goodreau et al., 2008), or summaries of distributions of graph statistics (Snijders et al., 2006).

An ERGM specified through a joint distribution (3.1) involving a choice of statistics or parameters corresponding to graph topological features in the negpotential function (3.2) will be referred to here as a traditional ERGM. This specification requires an explicit identification of global network features thought to reflect important aspects of graph topology that have scientific interpretations. For example, the social network interpretation of the transitivity parameter is that friends of friends are more likely to also be friends. Thus, the strength of traditional ERGMs is the ability to describe the graph in terms of understandable global features.

To end this section, we mention that fitting ERGMs to realized networks has been demonstrated to be a difficult task, particularly to a network with a large number of nodes. The model can become degenerate, or place most of its probability on a few, disparate graphs, none of which resemble the observed network. A large amount of research has been devoted to identifying the cause of this behavior (Bhamidi et al., 2008; Handcock, 2003a; Park and Newman, 2004, 2005), recognizing when it has occurred (Schweinberger, 2011; Hunter et al., 2008a), , and proposing modifications to the ERGM to avoid the issue (Snijders et al., 2006; Robins et al., 2007; Hunter, 2007). One hypothesized cause of the behavior are large and growing neighborhoods (Schweinberger and Handcock, 2012; Pattison and Robins, 2002), which lead to the local dependence dominating the global structure. LSGMs are not immune to this behavior, although increased interpretability of the dependence through controlled neighborhoods and

saturated graphs can permit an easier identification of when such degeneracy will occur, which we explain in the following section.

3.3 Local Structure Graph Model (LSGM)

Local structure graph models (LSGMs) are a new class of graph models, having a global or joint distribution defined in terms of interpretable and controllable local dependence structures. Two defining characteristics of LSGMs are the specification, for each potential edge marker $\mathbf{s}_i$, of a full conditional distribution, $\Pr(y(\mathbf{s}_i)|y(\mathbf{s}_j); j \neq i)$, for the probability of edge presence or absence ($y(\mathbf{s}_i) = 1$ or 0) and a neighborhood, $N_i = \{\mathbf{s}_j : \mathbf{s}_j \text{ is a neighbor of } \mathbf{s}_i\}$ consisting of graph edges which are "local" to $\mathbf{s}_i$. These two features, together with an assumption of Markov dependence induce a direct functional dependence between graph edges defined to be neighbors,

$$\Pr(y(\mathbf{s}_i)|y(\mathbf{s}_j); j \neq i) = \Pr(y(\mathbf{s}_i)|y(\mathbf{s}_j); \mathbf{s}_j \in N_i)$$

This allows the probability of the presence of an edge to be dependent upon the outcomes $y(\mathbf{s}_j)$ of its neighboring edges, $\mathbf{s}_j \in N_i$.

The LSGM can be motivated by a Markov Random Field (MRF) model defined on graph edges. MRF models are commonly encountered in the analysis of spatial data, where effects of spatial dependence are specified conditionally on spatial location information. Intuitively, a response at a particular spatial site might be most heavily influenced by those sites which are spatially neighboring. Through the definition of neighborhoods and conditional specification, a MRF for spatial data allows dependence to be defined through specification of a local structure.

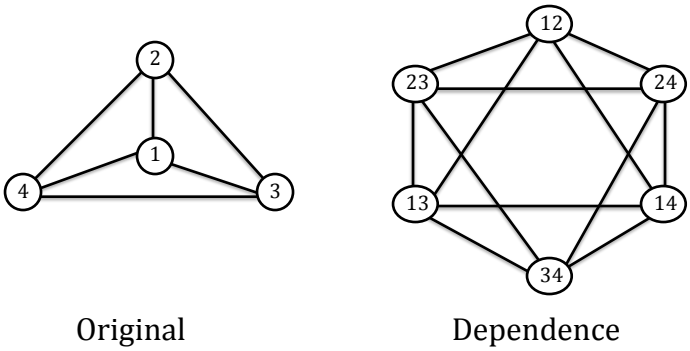
Most common applications of binary MRF models to spatial data can be associated with undirected graphs (Kaiser and Cressie, 2000). In these spatial problems, graph nodes correspond to binary random variables and edges connect nodes which are neighbors. To apply a MRF model to LSGMs, potential edges in the original graph become random variables (locations) in the MRF and neighborhoods are composed of sets of edges that are "near" each other according to some metric. Nodes of the original graph do not appear explicitly in the binary MRF model other than through their role in the edge markers.

To make this connection clearer, consider the neighborhood structure of a LSGM as represented through a dependence graph (see also Frank and Strauss, 1986). Each node in the dependence graph corresponds to a potential edge in the original graph where a connection in the dependence graph indicates the corresponding random variables are conditionally dependent. Two example networks and dependence structures with resulting MRF dependence graphs are shown in Figure 3.1. The first example appeared in Frank and Strauss (1986) and demonstrates the Markovian dependence as two edges are conditionally dependent if they are incident, or share a node. The second example demonstrates the potential flexibility in the definition of a neighborhood for edges. For this dependence structure, two edges are conditionally dependent if they connect the same number of red, or odd-numbered nodes. The dependence graph here is composed of three disconnected, yet internally fully connected, subgraphs of edges that join the same number of red nodes. In other words, because the nodes of the dependence graph represent potential edge occurrences as random variables in the original graph, the LSGM is placing a MRF on the nodes of the dependence graph.

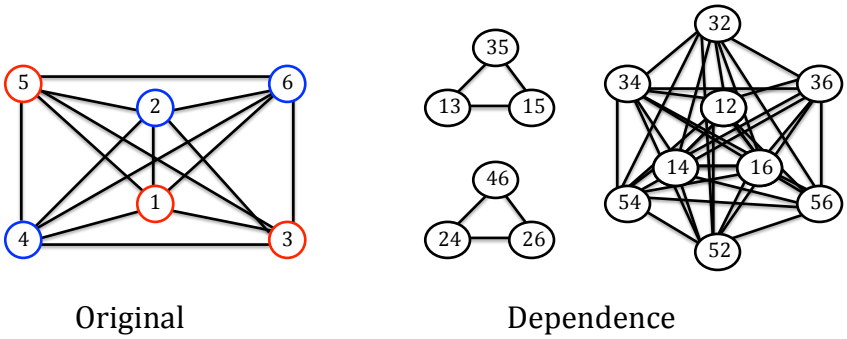
The idea of a neighborhood in network analysis has also been used elsewhere, although the definition of a neighborhood has not been consistent. Within LSGMs, a neighborhood defines edges which are conditionally dependent. Neighborhoods are often overlapping and a neighborhood is defined for each potential edge in the network. Our use of the term "neighborhood" (in connection to MRFs and LSGMs) differs from other common uses of this term in network analysis involving block models or community detection (Schweinberger and Handcock, 2012), where a "neighborhood" often implies a partitioning set of the nodes of the network.

3.3.1 Specification

To formulate a LSGM, one must specify the form of the conditional distributions and a neighborhood or dependence structure. For simple networks, the goal is to model the presence or absence of edges and thus the conditional distributions are binary, as with the initial LSGM described next. A binary conditional distribution expressed in exponential family form is given



(a) Incidence definition of dependence. Image recreated from Frank and Strauss (1986)



(b) Two edges are conditionally dependent if they connect the same number of red, or odd numbered, nodes.

Figure 3.1 Two example networks and dependence structures with resulting dependence graphs. The nodes of the dependence graph corresponds to the edges of the original graph. An edge in the dependence graph indicates conditional dependence between the two random variables.

by

$$\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i)) = \exp\left[y(\mathbf{s}_i)A_i(\mathbf{y}(N_i)) - B_i(\mathbf{y}(N_i))\right], \quad y(\mathbf{s}_i) = 0, 1 \tag{3.4}$$

where A_i is a natural parameter function and $B_i = \log[1 + \exp(A_i(\mathbf{y}(N_i)))]$. In (3.4), $\mathbf{y}(N_i)$ represents values of the binary random variables (here edges) in the neighborhood of $y(\mathbf{s}_i)$; note $y(\mathbf{s}_i) = 1$ indicates edge occurrence at the marker $\mathbf{s}_i$ of a potential edge. Dependence among random variables is modeled through the natural parameter function, A_i, and a function B_i of A_i. For binary conditionals, a form of the natural parameter function is

$$A_i(\mathbf{y}(N_i)) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i} \eta_{ij}[y(\mathbf{s}_j) - \kappa_j] \tag{3.5}$$

With m denoting the number of edge markers or total number of possible edges in the network, the sets of parameters, $\{\kappa_i : i = 1, \ldots, m\}$ and $\{\eta_{ij} : i = 1, \ldots, m; \mathbf{s}_j \in N_i\}$, represent global and local structure features of the network model, respectively, and will be discussed in detail in Section 3.3.2. The parameterization of the natural parameter function in (3.5) involves centering by global parameters κ_j, as introduced by Caragea and Kaiser (2009) and Kaiser et al. (2012). This centered parameterization has been shown to separate global from local structure in (3.4)–(3.5) leading to increased interpretation of all model parameters for reasonable amounts of statistical dependence.

The specification of any collection of full conditional distributions does *not* necessarily lead to a valid joint distribution on $\mathbf{Y}$, thus certain conditions must be satisfied. In formulating MRF models through conditional distributions, Kaiser and Cressie (2000) detail requirements for a joint distribution to exist with full conditionals matching those specified. In LSGM construction (3.4), a sufficient condition is that summation term in the natural parameter function (3.5) be symmetric for any pair of edges, implying that neighborhoods and η-parameters must be symmetric, i.e., $\mathbf{s}_i \in N_j$ implies $\mathbf{s}_j \in N_i$ and $\eta_{ij} = \eta_{ji}$. If the joint distribution exists, it is uniquely determined by the set of conditionals (Arnold and Press, 1989).

Section 3.2 discussed how an ERGM is defined by specifying global topological graph features, or parameters and statistics, to include in the negpotential function (3.2). Specification of the negpotential function is equivalent to the specification of a joint distribution (3.1). Graph

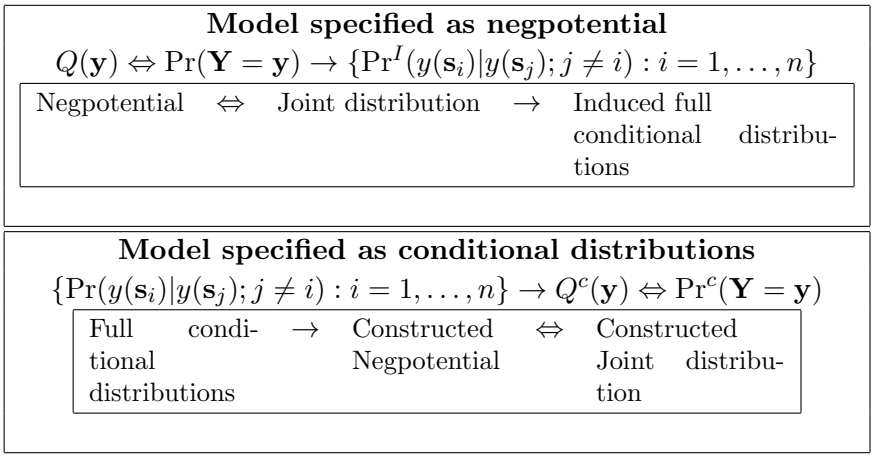


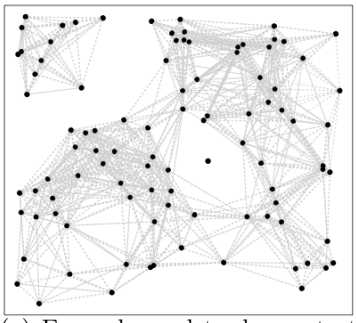
Figure 3.2 Relationship between the negpotential, joint distribution, and full conditional distributions when either the model is specified as the negpotnetial or full conditionals.

features chosen to be included in an ERGM implies a set of induced full conditional distributions, but, because an ERGM focuses on global network features, these conditional distributions are not directly modeled or often even identified. In contrast, LSGMs are defined by specifying a set of full conditional distributions which leads to a constructed negpotential function and thus joint distribution. This relationship between the two different methods of model specification is demonstrated in Figure 3.2. Using the binary conditionals from (3.5), a constructed negpotential function for a LSGM can be shown to be (Kaiser et al., 2012)

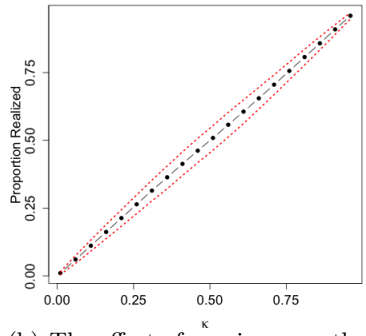
$$Q^C(\mathbf{y}) = \sum_{i=1}^{n} \left[\log\left(\frac{\kappa_i}{1-\kappa_i}\right) - \sum_{\mathbf{s}_j \in N_i} \eta_{ij}\kappa_j \right] y(\mathbf{s}_i) + \sum_{i=1}^{n} \sum_{\mathbf{s}_j \in N_i} \eta_{ij} y(\mathbf{s}_i) y(\mathbf{s}_j) \tag{3.6}$$

determining the joint distribution (3.1) for $\mathbf{Y}$ under these conditionals; above the superscript C denotes a negpotential Q^C constructed from full conditionals, e.g. (3.4), in contrast to a direct negpotential formulation (3.2). The functional form (3.6) implies a LSGM here can be represented as an ERGM (3.1) with Markovian dependence, and thus our proposed approach provides an alternate specification of a type of ERGM.

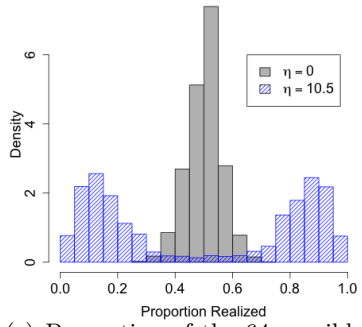
Network model features can generally be divided into those that affect the global structure and those that affect the local structure of random graphs (Schweinberger and Handcock, 2012; Goodreau, 2007). The global structure can be defined through patterns prevalent in the overall network, such as density. Features that allow for departures from the global structure at a local



(a) Example used to demonstrate effect of model parameters. Gray lines indicate all possible edges.



(b) The effect of varying κ on the proportion of possible edges realized. Points represent the median of 10,000 simulations and red, dashed lines 95% envelopes.



(c) Proportion of the 64 possible edges in the disconnected clump of the northwest corner realized for $\kappa = 0.5$ in 10,000 simulations obtained from two values of η.

Figure 3.3 Example network and a demonstration of the effect of model parameters.

level would be classified as local structure. An example of the local structure is transitivity, or tendency towards the closure of individual triangles. By specifying LSGMs through conditional distributions, we are able to directly model such local structure.

3.3.2 Model Parameters

Recall that, in LSGMs (3.4), two sets of parameters control the natural parameter function (3.5), $\{\kappa_i : i = 1, \ldots, m\}$ and $\{\eta_{ij} : i = 1, \ldots, m; \mathbf{s}_j \in N_i\}$. In its most general form, this model could allow for a different κ_i for every potential edge $y(\mathbf{s}_i)$ and a different η_{ij}, with $\eta_{ij} = \eta_{ji}$, for every $\mathbf{s}_j \in N_i$. However, restrictions are typically placed on these sets of parameters for model identifiability. The effect of the model parameters will be demonstrated for the simplest case where $\kappa_i = \kappa$ and $\eta_{ij} = \eta$ for every i, j in the example network displayed in the first panel of Figure 3.3. The network consists of 97 nodes and 824 possible edges, where only those pairs of nodes connected in Figure 3.3 are assigned a random variable for the potential occurrence of an edge and thus indicate those network edges with a positive probability of being realized.

Large-scale structure in (3.4)-(3.5) is represented by the first parameter to be discussed, $\kappa \in (0, 1)$. The parameter κ controls the density or proportion of realized edge variables in the overall network and is interpreted as the marginal probability a randomly chosen potential edge will be realized, $y(\mathbf{s}_i) = 1$. As a demonstration of the effect of κ, 10,000 networks

were simulated for 20 values of κ and a fixed $\eta = 5$. Due to the conditional specification, a network from a LSGM is naturally simulated with a Gibbs sampler where each potential edge is sampled from its conditional, $\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i))$, in turn. (Given randomly initialized values for all edges, the Gibbs sampler was run with a burn-in of 10,000 complete graph iterations after which sample graphs were retained from subsequent rounds of 500 iterations for thinning.) For the retained simulations, the proportion of realized edges out of those possible was computed for each graph. The second panel of Figure 3.3 plots the median proportion of realized edges for each κ against κ and red, dashed lines enclose 95% of the simulated proportions. A strong, monotonic relationship exists between κ and the proportion of realized edges with little variability, especially towards the boundaries of the parameter space.

More subtle is the effect on local structure determined by the dependence parameter, η. This parameter quantifies the strength of dependence between neighboring edges and thus controls the extent to which sets of edges exhibit a neighboring effect or behave independently. When $\eta = 0$, the summation term in the natural parameter function (3.5) that incorporates the value of neighboring edges is absent, i.e., $A_i(\mathbf{y}(N_i)) = \log\left(\frac{\kappa}{1-\kappa}\right)$, so that each edge formation consequently occurs according to an independent Bernoulli trial with success probability κ (the conditional probability of edge realization is equivalent to the marginal one, as expected under independence). In contrast, larger values of η induce neighbor effects on edge probabilities which can lead to groups of edges behaving in the same manner, e.g., all realized or all not realized. To illustrate, we again simulated 10,000 networks for each of two LSGMs: $\eta = 0$ and $\eta = 10.5$, both with a fixed $\kappa = 0.5$. Now considering the 64 possible edges in the disconnected northwest clump of the example network (Figure 3.3), we computed the proportion of realized edges among this local subset from each simulation run and the last panel of Figure 3.3 displays a histogram of these proportions across the 10,000 simulations. When $\eta = 0$, edge probabilities are unaffected by the rest of the network resulting in a distribution of proportions which are symmetric and centered at κ, as displayed in the gray, solid histogram of Figure 3.3. Few simulations resulted in less than 40% or more than 60% of the edges in this northwest subset being realized. However, for the larger dependence parameter value, $\eta = 10.5$, an induced dependence between neighboring edges is clear. Neighboring edges tended to behave in a group

fashion, with edges among this northwest subgroup either mostly all present or mostly all absent, resulting in a bimodal distribution of proportions, as displayed by the blue, dashed histogram of Figure 3.3. Note that the histogram for this strong dependence scenario is still centered at the expected global proportion of realized edges, $\kappa = 0.5$. That is, even when the local dependence is strong, the marginal mean κ in LSGMs is preserved over multiple simulations, due to the centered parameterization of the natural parameter function (3.5).

Additional modeling of the local dependence parameter is often necessary. In application of a LSGM, we recommend that this term be adjusted to account for unequal neighborhood sizes. It is common in spatial statistics for neighborhoods of random variables in a MRF to be similar in size, such as occurs, for example, with a four-nearest neighbor structure for a regular spatial lattice. However, neighborhoods for potential edges in LSGMs will often not result in equally-sized neighborhoods (see Figure 3.8 in Section 3.4 for an example). To allow the summation term in the natural parameter function (3.5) to have a uniform effect on edges of varying neighborhood size, we modify dependence parameters as

$$\eta_{ij} = \frac{\eta}{|N_i| + |N_j|} \tag{3.7}$$

where $|N_k|$ represents the size of the neighborhood of edge $y(\mathbf{s}_k)$. The summation of neighborhood sizes in the denominator of (3.7) assures that $\eta_{ij} = \eta_{ji}$, guaranteeing the identification of a joint distribution through construction of a negpotential function (Kaiser and Cressie, 2000).

A practical parameter space for $\eta \in \mathbb{R}$ is not as well defined as the large-scale parameter, $\kappa \in (0, 1)$. When the local structure of the model overwhelms the global structure, e.g., $|\eta|$ is "too large" compared to κ, the model will become degenerate and place most of its probability on unrealistic network realizations. As a demonstration, the proportion of realized edges in 10,000 simulations of the example network for a LSGM with parameter values $\kappa = 0.5$ and $\eta = 35$ is shown in Figure 3.4. Almost all edges are realized in all simulations, as the model places most of its probability on the nearly complete graph. This behavior has been recognized for the ERGM and, more generally, in a class of models for interactive systems (Strauss, 1986), and is similar to long-range dependence observed in the Ising model (Snijders, 2002). In the ERGM context, large and growing neighborhoods have been identified as a potential

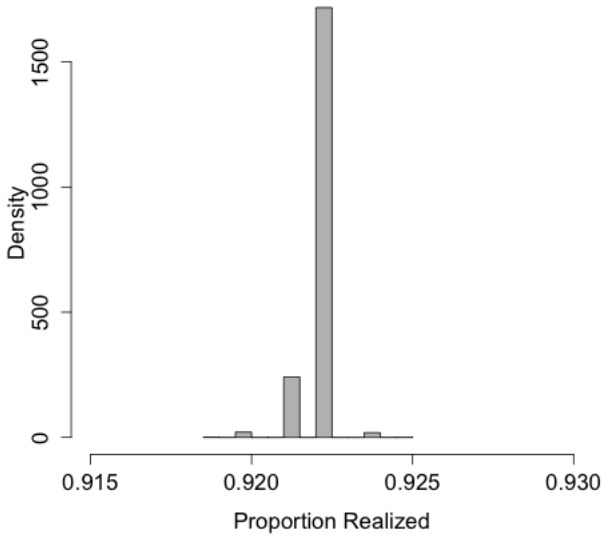


Figure 3.4 Proportion of realized edges in 10,000 simulations when $\kappa = 0.5$ and $\eta = 35$. The proportion realized does not correspond to the large-scale parameter, $\kappa = 0.5$. This is an example of an area of the parameter space where the model is degenerate.

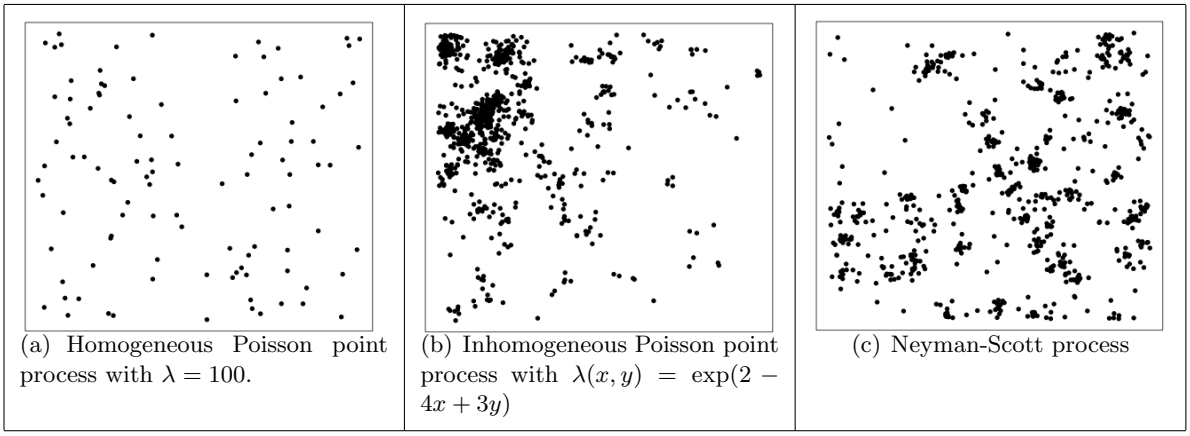
cause of degenerate model behavior (Schweinberger and Handcock, 2012). A large dependence parameter in LSGMs produces the same essential effect of model degeneracy as having overly large neighborhoods in the ERGM. Both result in summation terms in a negpotential function (e.g. (3.3), (3.5)) that dominate terms for marginal probability, which undermines any concept of dependence in the model (as a departure from independence), and thereby ruins the overall model. As a further complication related to similar degeneracy issues in ERGMs, the values for dependence parameters which are inappropriately large in a LSGM, leading to degeneracy, can change between data applications. A recommendation from Kaiser et al. (2012) is to simulate from the fitted LSGM to assure that the simulations appear reasonable given an observed network. Further work in this area is a topic of ongoing research, but the structures of clearly defined neighborhoods in a LSGM can help in diagnosing and treating degeneracy issues related to edge dependence.

3.3.3 Additional Features

An important issue in formulating LSGMs is how to define meaningful neighborhoods which capture an appropriate dependence structure. Two additional modeling features can be used to assist with this choice, a potentially latent spatial location of nodes (for defining neighborhoods given relevant spatial information) and a saturated graph (for restricting the total number of graph edges).

Recall that a LSGM, with its conditional specification and explicit neighborhood definition, incorporates ideas from the MRF model, a common tool used to analyze geo-referenced data. If the nodes of the network have an observed spatial location, such as the location of the buses in the electric power grid (Watts and Strogatz, 1998), the routers of the Internet (Neumayer and Modiano, 2010), or the formation site of tornadoes, LSGMs provide a natural way to incorporate this spatial information, particularly in formulating geographically defined neighborhoods for edges. Networks for which nodes do not have spatial locations can also be modeled with a LSGM. One option is to impose a latent spatial structure, and such types of spatial locations for nodes could be applied in defining neighborhoods. As an illustration, three example point processes and the resulting node placements are displayed in Figure 3.5. In this LSGM formulation, latent node locations could potentially be estimated iteratively as a step in a Gibbs sampler. As another example of imposing a spatial location for the nodes, a latent variable on the nodes might be imposed based on auxiliary information. That is, nodal covariate information could potentially be incorporated to define spatial locations for nodes in some unobserved "social space" (cf. Hoff et al., 2002; Handcock et al., 2007). However, to avoid introducing additional complicating factors, the application of latent point processes will not be considered in the current work; rather, the tornado example presented in Section 3.4 illustrates how spatial information may be incorporated to formulate a LSGM.

A saturated graph is a second additional LSGM feature that can assist in the specification of meaningful and useful neighborhoods for a network. A saturated graph is defined as those network edges having a positive probability of being realized, so that a saturated graph represents the maximal network realization under consideration. The network displayed in the first



(a) Homogeneous Poisson point process with $\lambda = 100$.

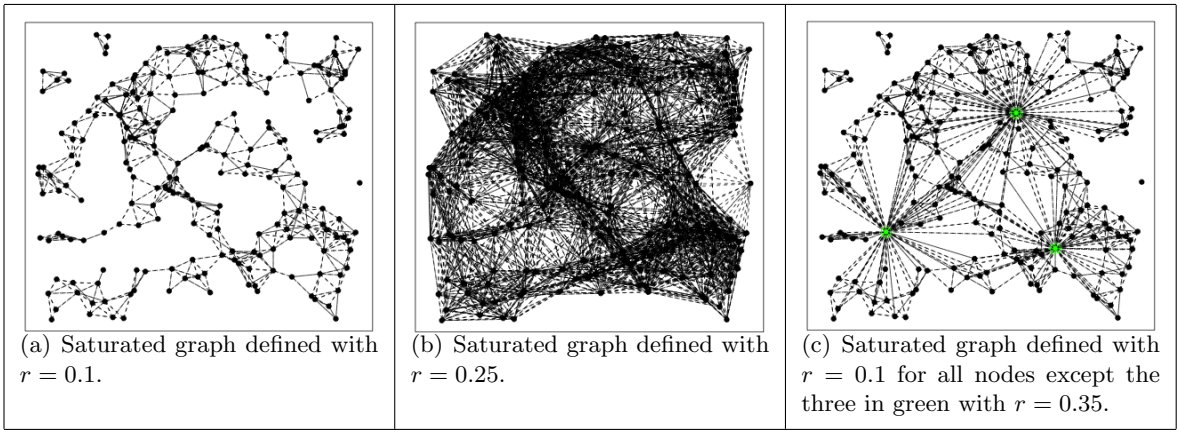
(b) Inhomogeneous Poisson point process with $\lambda(x, y) = \exp(2 - 4x + 3y)$

(c) Neyman-Scott process

Figure 3.5 Examples of random node placements through different point processes.

panel of Figure 3.3 is an example of a saturated graph. In some applications, it is reasonable to impose some types of cutoffs in potential edge formations, to produce a meaningful saturated graph for n nodes that is significantly smaller than a graph allowing $\binom{n}{2}$ edges. For instance, Sensor-Actuator Networks (SAN) have a common transmission range which is the maximum distance possible between two connected nodes (Onat and Stojmenovic, 2007) and, in biological networks, growth factors and diffusible signaling concentration decrease as a function of distance making "long distance" edges improbable (Sporns et al., 2004).

For networks composed of nodes with an observed or latent spatial setting, an intuitive approach for defining saturated graphs is to use a method similar to the formation of a unit disk graph (Kuhn et al., 2004). Given a radius, r, an edge between two nodes within distance r will be defined to have positive probability of being realized. To illustrate, three example saturated graphs with consistent node locations on the unit square are displayed in Figure 3.6. Radius size is held constant at $r = 0.1$ and $r = 0.25$, respectively, for all nodes in the first two panels. Smaller radius size leads to a graph that is not completely connected with two clusters of nodes disconnected from the majority and one isolated node. In contrast, the resulting saturated graph from the larger radius size is completely connected with no isolated nodes. Additionally in this manner, hubs of nodes could be permitted and modeled by varying the radius size between nodes. The saturated graph displayed in the rightmost panel of Figure 3.6



(a) Saturated graph defined with $r = 0.1$.

(b) Saturated graph defined with $r = 0.25$.

(c) Saturated graph defined with $r = 0.1$ for all nodes except the three in green with $r = 0.35$.

Figure 3.6 Examples of saturated graph on same set of nodes for various radius sizes.

was created with a radius of $r = 0.1$ for all except the three nodes highlighted in green which had radius $r = 0.35$.

One advantage to imposing a saturated graph is a decrease in the number of random variables to be modeled. As an illustration consider the three saturated graphs of Figure 3.6, which contain 213 nodes. The small radius of $r = 0.10$ results in 668 possible edges. When the radius is increased to $r = 0.25$, the number of possible edges jumps to 3,467; the combination of radius sizes results in 873 possible edges. Without a saturated graph, edges could form between all pairs of nodes which would result in $\binom{213}{2} = 22,578$ random variables to model. In a small example this may be plausible. However, the direction of current research is to analyze networks with a large number of nodes (Fienberg, 2012) so that modeling an edge between all pairs of nodes can be computationally prohibitive and perhaps physically unreasonable. Defining a saturated graph based on contextual information for the problem under consideration can be a beneficial modeling strategy.

As alluded to in Section 3.2, a further consequence of the saturated graph is reasonably sized neighborhoods. Consider an incidence definition of dependence, where two edges which share a node are conditionally dependent in the example network of Figure 3.6. In the absence of a saturated graph, each of the 22,578 edge random variables would be dependent upon $2(213 - 2) = 422$ neighbors, and thus each summation in the natural parameter (3.5) would include 422 terms. For the same incidence dependence structure, the use of a saturated graph

allows the neighborhood size to vary and depend on the number of "nearby" edges. The average neighborhood size for edges in the first panel of Figure 3.6 is 12.5, 68 for the second panel, and for the final panel each edge is dependent upon 30.67 neighbors, on average. Thus, the use of a saturated graph not only decreases computational time, but also can alleviate overly large neighborhood sizes, which was identified by Schweinberger and Handcock (2012) as a source of model degeneracy.

3.4 Application

In this section we formulate a LSGM for a network constructed from recorded tornadoes which originated within the state of Arkansas during April, 2011. Details of the tornadoes were obtained from National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Center *Storm Data* severe weather report database. The simplest LSGM with one global paramster, κ, and one local parameter, η, in (3.5) is fit to the Arkansas tornado network. Two model assessment techniques are presented to compare the fit of the LSGM to that of an independence model with only the global structure parameter, κ. Results indicate that this tornado network does exhibit a significant amount of local dependence and that the LSGM can more adequately capture and interpret this feature.

3.4.1 The Network

The nodes of the network represent the 59 tornadoes with a documented starting longitude and latitude that originated within Arkansas during April, 2011. Locations of nodes are a tornado's observed point of origin. Edges connect two tornadoes which belong to the same storm event (Agee et al., 1976). For successively occurring tornadoes to be within the same event, both must appear within two hours of each other and could have plausibly been produced by the same storm system. Thirteen storm events were identified to have produced at least one tornado with node location and storm event information displayed in Figure 3.7.

A saturated graph with a radius of $r = 80$ kilometers (cf. Section 3.3.3) was used for the Arkansas tornado network. The radius value was motivated by the fact that thunderstorms can travel upwards of 80 kilometers per hour. The saturated graph results in 292 possible edges,

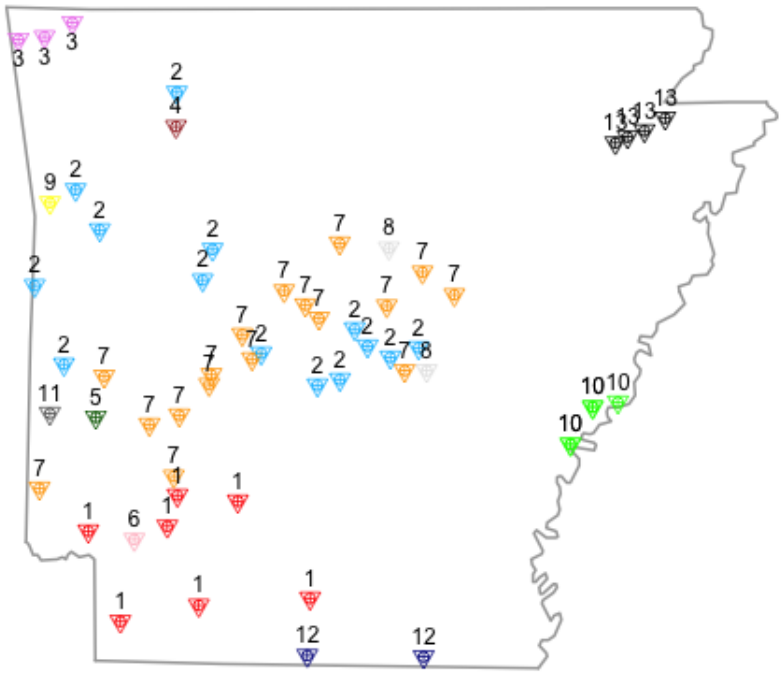


Figure 3.7 Nodes of the Arkansas tornado network defined by tornadoes that originated in Arkansas during April, 2011. Color and numbers correspond to the event in which the tornado occurred.

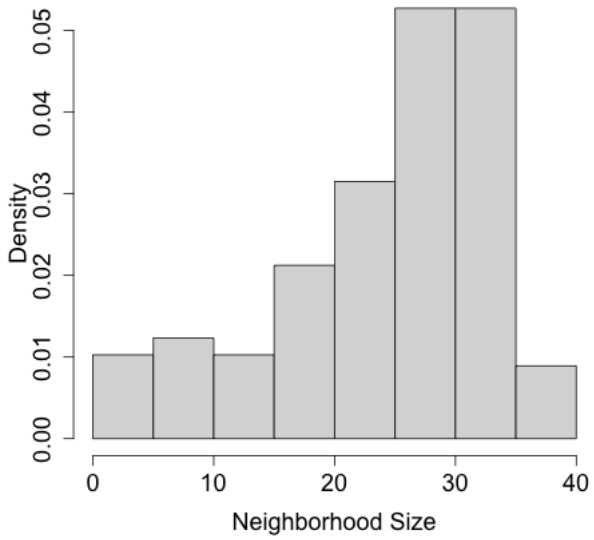


Figure 3.8 Neighborhood sizes when a saturated graph of $r = 80$ kilometers is used in the analysis of the Arkansas tornado network.

and thus requires modeling this many edge random variables. Using the incidence definition of dependence (two potential edges are conditionally dependent if they share a common node), the saturated graph leads to the neighborhood sizes shown in Figure 3.8. Restricting neighborhoods is justified by the application. A edge that connects two tornadoes in an area of the state where few tornadoes occurred, e.g., event 3 in Figure 3.7, is not dependent upon edges that occur in another part of the state where tornadoes are more likely, e.g., event 7.

3.4.2 The Fit of the LSGM

Here we consider the LSGM with a single marginal mean, κ, and single dependence parameter, η, for the Arkansas tornado network. The dependence parameter is adjusted to account for unequal neighborhood sizes as in (3.7). Point estimates of the model parameters are obtained through a maximization of the log pseudo-likelihood (PL), the summation of the log of the conditional distributions,

$$\log \mathrm{PL} = \sum_i \{y(\mathbf{s}_i) \log[p_i(N_i)] + (1 - y(\mathbf{s}_i)) \log[1 - p_i(N_i)]\}$$

Table 3.1 Point estimates and 90% percentile parametric bootstrap interval estimates for the LSGM and independence model fits to the Arkansas tornado network.

	$\hat{\kappa}$	$\hat{\eta}$
LSGM	0.27 (0.15, 0.75)	8.60 (4.93, 11.07)
Independence	0.43 (0.38, 0.48)	—

where $p_i(N_i) = E_i(Y(\mathbf{s}_i)|\mathbf{y}(N_i))$ represents the conditional expectation for edge $Y(\mathbf{s}_i)$ given neighboring values $\mathbf{y}(N_i)$,

$$p_i(N_i) = \frac{\exp(A_i(N_i))}{1 + \exp(A_i(N_i))}$$

The PL function was introduced by Besag (1975) as an approximation to the "true" likelihood function, providing a fast and computationally tractable method for obtaining point estimates from conditional distributions. Estimates obtained by maximizing the PL function for a MRF have been shown to be generally consistent and asymptotically normal (Guyon, 1995). Interval estimates were obtained using parametric bootstrap percentile intervals (Davison and Hinkley, 1997, Chapter 5.3) with 10,000 bootstrap renditions of the network data. For each simulated network, parameter estimates were obtained by PL and 90% percentile bootstrap confidence intervals were calibrated from the 5th and 95th percentiles of the resulting empirical distributions of estimates. Point estimates and 90% confidence intervals are shown in the first row of Table 3.1. For comparison purposes, a maximum PL estimate and parametric bootstrap interval are obtained for the one parameter independence model fit to the tornado network with the dependence parameter η set to zero. The results of this fit are also shown in Table 3.1.

3.4.3 Model Assessment

Two methods of model comparison are used to compare the LSGM to the independence model. The first involves a simulation-based analog of the likelihood-ratio test and the second approach attempts to quantify the extent to which the LSGM is able to replicate local structure of the network.

Likelihood ratio tests are commonly used to compare two models with nested parameter spaces. Because the likelihood is known only up to a constant for the LSGM and the model involves a lack of independence among random variables, using a likelihood ratio test to compare the LSGM to the independence model would become complicated. However, an approximate approach based on simulation and PL can be used. It is desired to test the fit of the null model, here independence model, against the fit of the alternative model, or LSGM. The test statistic is given by the difference in the maximized log-PL for both methods, or

$$D = \log \mathrm{PL}(\mathrm{LSGM}) - \log \mathrm{PL}(\mathrm{Indep}) \tag{3.8}$$

In order to assess the significance of the test statistic, a reference distribution is constructed through simulation. This is done by fitting both the LSGM and independence model to each of 10,000 networks simulated from the null, independence model, where (3.8) is computed for each simulation, $D_h^*,\ h = 1, \ldots 10,000$. The p-value for assessing whether the LSGM is significantly better than the independence model is then computed as

$$\frac{1}{10000} \sum_{h=1}^{10000} I(D_h^* > D) \tag{3.9}$$

where $I(A)$ is the indicator function which takes the value 1 if event A holds and 0 otherwise. For the Arkansas tornado network the fit of the two models yields $D = 14.06$ with a p-value of 0.0016. Thus, it can be concluded based on this test that the LSGM fits the Arkansas tornado network significantly better than the independence model. This implies that there is a significant amount of local dependence in the tornado data.

Assessing the ability of the LSGM in capturing local properties in the tornado network is the focus of the second model assessment technique. A model could be said to exhibit this trait if simulations from the fitted model are able to recreate local features of the observed network. The feature of interest will be neighborhood homogeneity, or the proportion of neighboring edges which have the same value as a given edge. If the edge at marker $\mathbf{s}_i$ is not present, i.e., $y(\mathbf{s}_i) = 0$, the proportion of neighbors which assume the same value is computed as

$$q(\mathbf{s}_i) = \frac{1}{|N_i|} \sum_{\mathbf{s}_j \in N_i} [1 - y(\mathbf{s}_j)]$$

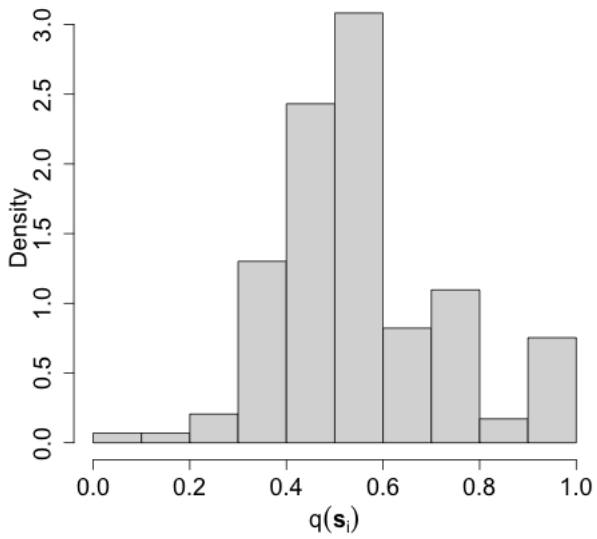


Figure 3.9 Proportions of neighbors assuming the same value as the random variable, $q(\mathbf{s}_i)$ for the Arkansas tornado network.

and if the edge at location $\mathbf{s}_i$ is realized, i.e., $y(\mathbf{s}_i) = 1$, the proportion is

$$q(\mathbf{s}_i) = \frac{1}{|N_i|} \sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j).$$

This results in an empirical distribution of proportions $\{q(\mathbf{s}_i), i = 1, \ldots, m\}$ with a proportion for each of the $m = 292$ potential edges in the Arkansas tornado network, as shown in Figure 3.9.

To demonstrate how well the LSGM captures this feature, consider the mean of the proportion of same neighbors, $\overline{q} = \frac{1}{m} \sum_{i=1}^{m} q(\mathbf{s}_i)$. For the Arkansas tornado network displayed in Figure 3.9 this is $\overline{q} = 0.561$. This average proportion can be computed for simulated networks from both models. Based on 10,000 simulations, a set of such proportions from each model, $\{\overline{q}^*_{\text{Indep},h}; h = 1, \ldots, 10,000\}$ from the independence model and $\{\overline{q}^*_{\text{LSGM},h}; h = 1, \ldots, 10,000\}$ from the LSGM, can be used as reference distributions to test the significance of the average proportion, $\overline{q} = 0.561$, from the observed network with p-values computed in a manner similar to (3.9). For the independence model, the p-value is 0.0002, while the p-value is 0.7481 for the LSGM. From this it can be concluded that the independence model is not able to capture

the local structure of the network as defined through neighborhood homogeneity but that the LSGM, with its neighborhood definition and dependence parameter, is able to recreate this feature.

The previous model assessment techniques indicate that the Arkansas tornado network exhibits a significant amount of local dependence that is appropriately captured by the LSGM. To better understand the nature of this local dependence in the LSGM, it is helpful to consider how the conditional probability of an edge (tornado siting) changes under the model as a function of neighboring outcomes. For example, consider an edge, $y(\mathbf{s}_i)$, with 20 neighbors, $|N_i| = 20$, where each of its neighbors also has 20 neighbors, $|N_j| = 20\ \forall \mathbf{s}_j \in N_i$. Under the fitted LSGM, the marginal expectation for this edge being realized is $\hat{\kappa} = 0.27$, regardless of the value of the neighbors. However, the conditional probability,$p_i(N_i)$, that this edge occurs, depends heavily on the number of realized neighboring edges. This relationship is plotted in Figure 3.10. When all neighboring edges of $y(\mathbf{s}_i)$ are absent, the conditional probability that $y(\mathbf{s}_i) = 1$ is only 0.10. The probability increases monotonically with the number of realized neighboring edges to $p_i(N_i) = 0.89$ when all neighboring edges are realized, i.e., $y(\mathbf{s}_j) = 1\ \forall \mathbf{s}_j \in N_i$.

3.5 Conclusions

The goal of this work is to introduce local structure graph models (LSGMs), a new class of models for network analysis, and to demonstrate its use with a simple application. Specification of a LSGM is achieved through conditional distributions which are functions of specified edge neighborhoods, or sets of conditionally dependent edges. An advantage of the LSGM approach is an explicit formulation of local dependence in the network, resulting in dependence which is interpretable and controlled by the modeler.

Behavior of LSGMs is controlled by two sets of parameters in a binary model for graph edges: parameters $\{\kappa_i; i = 1, \ldots, m\}$ which represent the global structure for the network model and control the marginal probabilities of edge realization in the network, and parameters $\{\eta_{ij}; i \neq j\}$, which capture the local model structure and can be interpreted as dependence parameters. If dependence parameters become too large, LSGMs can become degenerate, a common modeling

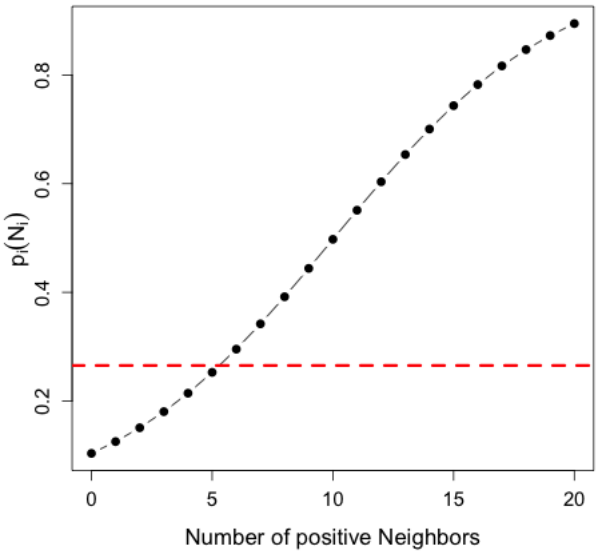


Figure 3.10 Number of positive neighbors against conditional expectation for a random variable with 20 neighbors. The red, dashed, vertical line represents the marginal expectation of $\hat{\kappa} = 0.27$.

consideration for models of interactive systems which encompasses ERGMs. However, because LSGMs are connected to edge neighborhoods through their specification, this aspect may help in formulating and diagnosing models which avoid model degeneracy through controlled, local dependence parameters; this is a topic of on-going investigation. Spatial location of nodes and a saturated graph are introduced to aid in avoiding model degeneracy. These features are not required to the specification of a LSGM, as the form of the conditional distributions and neighborhood structure is all that is necessary.

An extension to LSGMs is the inclusion of auxiliary information into either the global or local (dependence) structure. This can be accomplished through additional modeling of κ, η, or the neighborhoods. Explicit modeling of transitivity, or dependence between triples of random variables, will also require an additional extension. This is due to the fact that the constructed negpotential of a LSGM in (3.6) includes an assumption of pairwise-only dependence, where dependent sets of random variables of size greater than two are not directly modeled. Although this assumption is often appropriate for the common spatial application of a MRF model, it may be less suitable for the analysis of some networks.

CHAPTER 4. LOCAL STRUCTURE GRAPH MODELS WITH HIGHER-ORDER DEPENDENCE

Abstract

Local Structure Graph Models (LSGMs) provide a Markov Random Field (MRF) modeling approach for random graphs, whereby each edge in the graph has a specified conditional distribution, i.e., probability of edge occurrence, dependent on explicit neighborhoods, or subcollections of other graph edges, that define a conditional distribution. As a consequence of the conditional specification, LSGMs have the advantage of allowing direct control and separate interpretation of parameters influencing large-scale (e.g., marginal means) and small-scale (i.e., dependence) structures in a graph model. This is possible through so-called *centered parameterizations* of MRF models, which are applied in LSGMs. However, current technology for centered parameterizations in MRFs assumes pairwise-only dependence, meaning that dependence is modeled between pairs of random variables only. This creates limitations in specifying conditional distributions for graph edges in LSGMs. As a remedy, we extend the centered parameterization for MRFs to account for triples of dependent edges in LSGMs. We also explain and numerically illustrate the importance of centered parameterizations when interpreting model parameters and, using a MRF framework, we additionally show that common exponential random graph models induce conditional distributions without centered parameterizations and thereby have undesirable consequences in parameter interpretation compared to LSGMs. Centered parameterizations and their increased interpretation are particularly crucial when attribute/covariate information is included in a graph model. We illustrate these aspects for LSGMs with two network data examples, where attribute information is included in large and small-scale structures and where dependence between triples of dependent edges is explicitly modeled. This work hence advances the modeling of graph data in several important ways

related to conditional model specifications, state-of-the-art parameterizations and inclusions of higher-order dependence, and appropriate model incorporation of covariates.

4.1 Introduction

The study and application of network science has become increasingly popular within many fields of research. A network is composed of a set of nodes and the relations between them. In general, a network represents relational data, or data containing features that go beyond the information contained within individual nodes (Handcock et al., 2008). Complex patterns of connections and dependencies can be represented within the framework of a network and, due to this ability, networks can be applied to a diverse array of problems from many disciplines. Network science as a field is sizable, diverse, and rapidly growing.

Local structure graph models (LSGMs), introduced in Casleton et al. (2014b), are a novel approach to formulating network models using sets of conditional distributions and explicitly defined neighborhoods for random variables that represent edges in the network. The advantage of this network modeling approach is control over the local structures in the network based on the application of a Markov Random Field (MRF) model to the edges of a network. LSGMs can also be considered as an alternate method of specifying an exponential random graph model (ERGM) (Kolaczyk, 2009, p. 180). In contrast to LSGMs, traditional formulations of ERGMs specify a model for a network through a joint distribution, by identifying particular global topological graph features to be included as statistics in the log-linear term of the joint distribution. Frequently included features are edge density, transitivity, block effects, or covariate effects. Sets of dependent edge random variables and conditional distributions are induced by the terms included in the joint model, rather than explicitly defined as in a LSGM. Both the traditional formulations of ERGMs and LSGMs have joint distributions in Gibbsian form (Casleton et al., 2014b), but these joint distributions must be constructed for a LSGM from the set of specified conditional distributions (Kaiser and Cressie, 2000), which may be accomplished under certain conditions.

As one important contribution here, we aim to contrast models for random graphs from both LSGM and ERGM approaches using a framework of MRFs and conditional distributions,

which reveals that LSGMs can have useful benefits in parameter interpretation. Within the context of MRF models, the parameterization of conditional distributions has been shown to have an important effect on model parameter interpretation. Parameters which represent the large-scale model structures in the original parameterization of Besag (1974) (e.g., marginal means) can be influenced by small-scale model structures or the amount of statistical dependence between neighboring random variables (Caragea and Kaiser, 2009). This undesirable feature ruins the modeling intention and complicates the comparison of parameters between varying data sets. This aspect is also particularly troublesome when attribute information, such as covariate information on the nodes, is modeled. Alternatively, the recent centered parameterization of Caragea and Kaiser (2009), Hughes et al. (2011) and Kaiser et al. (2012) improves the interpretability of MRF models by separating the interpretation of the parameters which represent the large- and small-scale model structures. Hence, as an advantage of their specification of graph models through conditional distributions, LSGMs have the ability to directly build in this state-of-the-art centered parameterization. This parameterization also allows for the interpretable inclusion of nodal covariates into the conditional specification of a random graph model through the LSGM approach. On the other hand, ERGMs also correspond to a MRF model, but it will be shown that the conditional distributions induced by their joint specification correspond to the original, uncentered parameterization of MRF models. This implies that model parameters in ERGMs can be confounded in often unappreciated and undesirable ways which also has implications for model degeneracy.

As a related and further consideration here in modeling graphs, implementation of MRF models (particularly for spatial data) often assume pairwise-only dependence, so that dependence is explicitly modeled only between pairs of random variables. For valid MRF models, a necessary form of conditional distributions that allow for dependence between an arbitrary set of random variables is presented in Lee et al. (2001). However, due to its ubiquitousness in spatial applications, the centered parameterization was developed with this assumption and the notion of centered parameterizations has not been extended to higher-order dependence terms, which would be valuable for LSGMs with graph data. In this article we extend the centering ideas to the important situation within network analysis where the model explicitly

accounts for dependence between triples of random variables. This development applies to a LSGM, and more generally, to MRF models.

The remainder of the article is organized as follows. Section 4.2 describes the centered parameterization and contrasts it to the original parameterization of Besag (1974) within the context of a MRF model. Section 4.3 applies this parameterization to random graph models and examines the induced parameterization of an ERGM. Section 4.4 relaxes the pairwise-only dependence assumption in the MRF specification of LSGMs while maintaining a centered parameterization in the situation where dependence between triples of edges is modeled. In Section 4.5, a LSGM is used to incorporate auxiliary information in a network analysis of plant succession, and an example network constructed from a season of football games is used to demonstrate the inclusion of the higher-order terms in a LSGM. These examples illustrate different aspects in model development for networks, related to how large- or small-scale structures may be specified with conditional distributions, centered parameterizations, and attribute information. Section 4.6 concludes the paper.

4.2 Parameterization for MRF Models

A network, or graph, is defined by a set of n nodes and m edges. The networks of interest here are undirected and simple, with unweighted edges and no self-loops. To construct a random graph model, assign to each of the $m = \binom{n}{2}$ possible edges a binary random variable $Y(\mathbf{s}_i)$, where the marker $\mathbf{s}_i = \{c_i, r_i\}$ indicates the two nodes, c_i and r_i, that a potential edge would join. Edge values are collected into $\mathbf{Y}$, an $n \times n$ adjacency matrix, and each entry designates the presence, $y(\mathbf{s}_i) = 1$, or absence, $y(\mathbf{s}_i) = 0$, of an edge between each node pair in the graph. For undirected, simple networks, $\mathbf{Y}$ will be symmetric with $\text{diag}(\mathbf{Y}) = 0$. A realization of the network will be represented as $\mathbf{y}$. Exogenous information, such as covariate information on the nodes or block effects, can be associated with the marker $\mathbf{s}_i$ and will be designated as a possible vector-valued $\mathbf{x}(\mathbf{s}_i)$.

4.2.1 Original Parameterization

The development of LSGMs for simple networks is motivated by a binary MRF model, also referred to as the auto-logistic model by Besag (1974). Two defining characteristics are the specification, for each potential edge denoted by a marker $\mathbf{s}_i$, of a full conditional distribution, $\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\{\mathbf{y}(\mathbf{s}_j); j \neq i\})$, for the probability of presence or absence of a particular edge ($y(\mathbf{s}_i) = 1$ or 0) and a neighborhood, $N_i = \{\mathbf{s}_j : \mathbf{s}_j \text{ is a neighbor of } \mathbf{s}_i\}$ consisting of edges which are "local" to $\mathbf{s}_i$. Let $\mathbf{y}(N_i) = \{y(\mathbf{s}_j) : \mathbf{s}_j \in N_i\}$ represent the values of the neighbors of $Y(\mathbf{s}_i)$. A Markov assumption results in the full conditional distributions being functionally dependent only upon the neighboring random variables, $\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\{\mathbf{y}(\mathbf{s}_j); j \neq i\}) = \Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i))$.

Consider the conditional binary probability function written in exponential family form as

$$\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i)) = \exp[y(\mathbf{s}_i)A_i\{\mathbf{y}(N_i)\} - B_i\{\mathbf{y}(N_i)\}], \quad y(\mathbf{s}_i) = 0, 1 \tag{4.1}$$

where $A_i\{\mathbf{y}(N_i)\}$ is referred to as the natural parameter function and $B_i\{\mathbf{y}(N_i)\}$ is a function of $A_i\{\mathbf{y}(N_i)\}$, which for an auto-logistic model is $B_i\{\mathbf{y}(N_i)\} = \log[1 + \exp(A_i\{\mathbf{y}(N_i)\})]$. The parameters of the model are contained in the natural parameter function, $A_i\{\mathbf{y}(N_i)\}$, and thus the parameterization of this function is of interest. In the original auto-logistic model of Besag (1974), the natural parameter function takes the form, for $i = 1, \ldots, m$,

$$A_i\{\mathbf{y}(N_i)\} = \alpha_i + \sum_{\mathbf{s}_j \in N_i} \eta_{ij} y(\mathbf{s}_j) \tag{4.2}$$

where $\{\alpha_i\}$ are leading constants and the $\{\eta_{ij}\}$ are dependence parameters between pairs of random variables. This parameterization will be referred to as the original or uncentered parameterization.

An undesirable feature of the original parameterization is that, while $\exp(\alpha_i)/(1+\exp(\alpha_i))$ is the expected value of $Y(\mathbf{s}_i)$ under an independence model (all $\eta_{ij} = 0$), it is not the marginal expected value of $Y(\mathbf{s}_i)$ under a model that includes dependence. This makes interpretation of estimated parameter values difficult, especially in models that include covariate information (Caragea and Kaiser, 2009).

This confounding of parameter interpretation can also contribute to complications with respect to model degeneracy. Model degeneracy occurs when a model places a large amount of probability on a small subset of the theoretically possible realizations, which may not resemble an observed set of data. This failure has been widely studied for ERGMs, though it has also been recognized in a more general class of models for interactive systems (Strauss, 1986) and has been associated with long-range dependence in Ising models (Snijders, 2002). Most of the work on diagnosing the cause of model degeneracy within ERGMs has identified a parameter space issue (see, for example Handcock, 2003a; Park and Newman, 2004, 2005; Rinaldo et al., 2009; Schweinberger, 2011). Thus, there are regions of the parameter space for which the model still theoretically exists, but does not perform well in practice. When parameter interpretation is confounded, identifying the offending regions of the parameter space becomes more difficult.

For later use and following Caragea and Kaiser (2009), let $p_i = \Pr(y(\mathbf{s}_i) = 1|\mathbf{y}(N_i)) = \Pr(y(\mathbf{s}_i)) = 1)$ under an independence model and $c_i = \Pr(y(\mathbf{s}_i) = 1|\mathbf{y}(N_i))$ in the presence of dependence. The log odds ratio in the presence of dependence relative to the independence model for the parameterization in (4.2) is

$$\log\left[\frac{c_i/(1-c_i)}{p_i/(1-p_i)}\right] = \sum_{\mathbf{s}_j \in N_i} \eta_{ij} y(\mathbf{s}_j), \quad i = 1, \ldots, m. \tag{4.3}$$

4.2.2 Centered Parameterization

To allow a more uniform interpretation of parameters across reasonable levels of statistical dependence, Caragea and Kaiser (2009) proposed a centered parameterization of the binary conditional distribution,

$$A_i\{\mathbf{y}(N_i)\} = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i} \eta_{ij}[y(\mathbf{s}_j) - \kappa_j], \quad i = 1, \ldots, m. \tag{4.4}$$

To compare this to the uncentered parameterization, from (4.4) the log odds ratio relative to the independence model was shown to be

$$\log\left[\frac{c_i/(1-c_i)}{p_i/(1-p_i)}\right] = \sum_{\mathbf{s}_j \in N_i} \eta_{ij}[y(\mathbf{s}_j) - \kappa_j] \tag{4.5}$$

which, unlike (4.3) for positive η_{ij} values, implies the odds that $y(\mathbf{s}_i) = 1$ increases if the number of realized neighbors is more than expected under the independence model and decreases if

this number is less than expected under independence. The advantage of the centered parameterization is the separating of large and small scale model structures, represented by the $\{\kappa_i\}$ and $\{\eta_{ij}\}$ sets of parameters, respectively. Model parameters can then be interpreted independently, which leads to a cleaner interpretation of auxiliary information, when available, and a better ability to identify the regions of the parameter space which will lead to degeneracy.

The ability of the centered parameterization to separate large and small scale model components has implications for the inclusion of auxiliary information in the form of covariates. The original development of MRF models Besag (1974) did not explicitly consider additional information in the form of covariates. Early extensions to the original work incorporated this information into the leading constant as $\alpha_i = \mathbf{x}(\mathbf{s}_i)^T\boldsymbol{\beta}$ in (4.2). For example, Gumpertz et al. (1997) incorporated soil variables into the leading constant for modeling the presence or absence of a disease in a field and Zhu et al. (2005) included time as a covariate in this term. But, both of these studies utilized the original parameterization, (4.2), and thus the interpretation of the effect of covariates is confounded with the strength of statistical dependence between pairs of neighboring random variables.

4.3 Parameterization for Random Graph Models

The previous section presented two parameterizations for the natural parameter function of a binary MRF model, given in expressions (4.2) and (4.4), respectively. The parameterization of both traditional ERGM and LSGM approaches to formulating network models can be examined within the context of binary MRFs. With its conditional specification, a LSGM is a more direct application of the binary MRF, but the conditional distributions that correspond to the joint specification of a traditional ERGM also define a MRF. In this section we determine the parameterization of the conditional distributions induced by the joint specification of an important case ERGM and show that these have the undesirable form of an uncentered parameterization (Section 4.2.1).

There are many possible forms of ERGMs defined by the many possible statistics to include in the joint distribution. To align this discussion with the pairwise-only dependence assumption, we will first consider a ERGM having only density and two-star effects. Although this model

is not currently widely used, it will be shown that the results generalize to other ERGM specifications. Assume that, in a given network of n nodes, all $m = \binom{n}{2}$ edges are possible and are thus assigned random variables.

The joint distribution of an ERGM with first order (density) and second order (two star) effects is

$$\Pr(\mathbf{Y} = \mathbf{y}) = c^{-1} \exp\left[\rho \sum_{i=1}^{m} y(\mathbf{s}_i) + \frac{\sigma}{2} \sum_{i=1}^{m} y(\mathbf{s}_i) \left(\sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j) \right) \right] \tag{4.6}$$

where c is a normalizing constant, ρ is a parameter related to density and σ is a parameter related to clustering, and the corresponding sums in (4.6) count the number of realized edges and 2-stars. For this specification, the dependence structure is incident, i.e., two edges which share a common node are dependent, so that neighborhoods can be determined from this. Designate the resulting neighborhood of a potential edge with marker $\mathbf{s}_i = \{c_i, r_i\}$ as

$$N_i = \{\mathbf{s}_j = (c_j, r_j) : c_j = c_i \text{ or } r_j = r_i\}, \quad i = 1, \ldots, m.$$

To count the number of two-stars realized in a particular graph, we multiply each pair of edge outcomes that share a node and sum the resulting products. If both edges are realized, the product will be 1; alternatively, if one or neither edge is realized the product will be 0. Thus, if a particular random variable is realized so that $y(\mathbf{s}_i) = 1$, then the count of the number of two-stars formed with that particular edge is $y(\mathbf{s}_i)S_i$ for

$$S_i = \sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j), \quad i = 1, \ldots, m.$$

.

Based on the neighborhoods N_i and sums S_i above, we may now determine the form of the conditional distributions implied by the joint specification of a traditional ERGM as in expression (4.6).

Proposition 4.3.1. *The full conditional distributions of an ERGM with density and two-star parameters are given by*

$$\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i) | \mathbf{y}(N_i)) = \frac{\exp[y(\mathbf{s}_i) A_i\{\mathbf{y}(N_i)\}]}{1 + \exp[A_i\{\mathbf{y}(N_i)\}]}, \quad y(\mathbf{s}_i) = 0, 1, \tag{4.7}$$

where $A_i\{\mathbf{y}(N_i)\} = \rho + \sigma S_i, i = 1, \ldots, m = \binom{n}{2}$.

A proof of the proposition is given in the appendix.

From Proposition 4.3.1, we have the result below establishing that conditional distributions following from a ERGM have the unattractive (uncentered) parameterization (4.2); see the appendix for a proof.

Corollary 4.3.2. *The implied conditional distribution of an ERGM with density and two-star parameters corresponds to an uncentered parameterization as in* (4.2)*.*

Proposition 4.3.1 and its corollary may be immediately extended to other ERGMs specifications containing exogenous information, such as attributes. Exogenous attributes do not depend on the structure of the graph and can be incorporated at the level of the individual nodes, edges, or as symmetric functions of nodal covariates, without affecting the implied dependence structure.

Exogenous attributes of nodes may be included through main effects which allow for a different parameter value based on the covariate value of the node (Goodreau, 2007). A similar example for pairs of nodes is assortative mixing, which attempts to capture the increased probability of edges to form between nodes within a particular attribute class (Goodreau et al., 2008). When this effect is uniform across attribute classes, it is referred to as uniform homophily, where differential homophily allows for a different parameter to describe this effect within the different attribute classes.

Regardless of the manner in which the exogenous information is included, the form of the parameter and statistic to be included in the joint distribution of this ERGM can be generalized. Assume there are p attribute effects to be included in the model. Further assume these effects will be included in a model having the endogenous effects of density and two-star. The resulting joint distribution is

$$\Pr(\mathbf{Y}=\mathbf{y}) = c^{-1}\exp\left[\rho\sum_{i=1}^{m} y(\mathbf{s}_i) + \frac{\sigma}{2}\sum_{i=1}^{m} y(\mathbf{s}_i)S_i + \sum_{j=1}^{p}\beta_j\sum_{i=1}^{m} y(\mathbf{s}_i)f(\mathbf{x}(\mathbf{s}_i))\right]$$

where β_j is the parameter associated with covariate j and $f(\mathbf{x}(\mathbf{s}_i))$ is some function of the covariate information associated with edge marker $\mathbf{s}_i = \{c_i, r_i\}$. For example, if a uniform homophily term for attribute j is included, the function will be $f(\mathbf{x}(\mathbf{s}_i)) = I(x_j(r_i) = x_j(c_i))$

where $x_j()$ denotes the value of the jth covariate for nodes r_i and c_i, i.e., functions of the marker information $\mathbf{x}(\mathbf{s}_i)$, and $I(\cdot)$ denotes the indicator function. The resulting natural parameter function of the implied conditional distributions from a ERGM with attribute information will include this information in the leading term α_i in (4.2). This results from the fact that the attribute information does not alter the dependence structure of the model. Attribute information in the leading term of the natural parameter function combined with an uncentered parameterization means the interpretation of the effect of the covariate is confounded with the amount of statistical dependence, represented here by the clustering term, σ.

Because models with a centered parameterization of the natural parameter function, such as the LSGM, have been shown to be able to separate the large and small scale model components, this has advantages for the inclusion of exogenous information as well. In addition, the flexibility of specifying the model through conditional distributions allows for this information to easily be included into either the large or small scale model component and, combined with the centered parameterization, allows the effects to be interpreted independently. As an example, consider a friendship network between school-aged children with covariate information of gender on the nodes of the network. This gender covariate information can be included in the large-scale model component. On the other hand, another example on plant succession (to be discussed in Section 4.5) demonstrates the inclusion of auxiliary information into the dependence between edges of the network, i.e., into small-scale model components.

4.4 Higher-Order Dependence

The original and centered parameterizations, (4.2) and (4.4), assume pairwise-only dependence. This implies that models with either parameterization allow for explicit modeling of the dependence between pairs of random variables, but do not account for any higher-order dependence. The assumption is common in spatial statistics, but network analysis may require inclusion of higher-order dependence to account for phenomena such as transitivity.

To incorporate higher-order dependence, consider first the negpotential function. Assume the network contains a finite number m of possible edges. Let the support of the joint distribution of $\mathbf{Y} = \{\mathbf{Y}(\mathbf{s}_i) : i = 1, \ldots, m\}$ be designated as Ω and assume there exists a $\mathbf{y}^* \in \Omega$ such

the that joint probability distribution is positive, $\Pr(\mathbf{Y} = \mathbf{y}^*) > 0$. Define the negpotential function to be

$$Q(\mathbf{y}) \equiv \log\left[\frac{\Pr(\mathbf{Y} = \mathbf{y})}{\Pr(\mathbf{Y} = \mathbf{y}^*)}\right], \quad \mathbf{y} \in \Omega.$$

Using the specific value $\mathbf{y}^* = \mathbf{0}$, Besag (1974) showed that the negpotential function may be expanded on Ω as

$$\begin{aligned} Q(\mathbf{y}) = &\sum_{1\le i\le m} H_i(y(\mathbf{s}_i)) + \sum_{1\le i<j\le m}\sum H_{i,j}(y(\mathbf{s}_i), y(\mathbf{s}_j)) \\ &+ \sum_{1\le i<j<k\le m}\sum\sum H_{i,j,k}(y(\mathbf{s}_i), y(\mathbf{s}_j), y(\mathbf{s}_k)) \\ &+ \dots \\ &+ H_{1,2,\dots,m}(y(\mathbf{s}_1), y(\mathbf{s}_2), \dots, y(\mathbf{s}_m)) \end{aligned} \tag{4.8}$$

and this was generalized to other values of $\mathbf{y}^* \in \Omega$ by Kaiser and Cressie (2000).

The Hammersley-Clifford theorem (Cressie, 1993, p. 417) gives that an H function in (4.8), $H_{i,j,\dots,g}$, is equal to zero unless the random variables at markers $\{\mathbf{s}_i, \mathbf{s}_j, \dots \mathbf{s}_g\}$ form a clique. A clique is defined as a single random variable or any set of random variables such that all random variables in the set are neighbors of every other random variable in the set. For example, a common neighborhood structure when random variables are arranged on a regular spatial lattice is the four-nearest neighbors. In this scenario, cliques exist only up to size two. Therefore, any H function which contains more than two random variables is zero by the Hammersly-Clifford theorem and thus, for this neighborhood structure, pairwise-only dependence is a consequence of this theorem. In contrast, with eight-nearest neighbors, cliques exist up to size four. However, the original formulation of the natural parameter function is often still used, and the summations that involve cliques of size three and four are assumed zero.

Conversely, the Hammersly-Clifford theorem can be used to identify the dependent sets of random variables when the model is specified as a joint distribution, such as for ERGMs. The ERGM considered in Section 4.3 with density and two-star parameters (4.6) was stated to induce an incidence definition of neighbors, so that edges which do not share a node are

conditionally independent. This was determined by examining the terms in the negpotential function and applying the Hammersly-Clifford Theorem.

As shown in Casleton et al. (2014b), a correspondence between ERGMs and LSGMs is demonstrated through the negpotential function. The joint distribution for both is that of a Gibbs distribution, which can be specified through the negpotential function, up to a constant. ERGMs directly model the joint distribution by specifying the negpotential function, but LSGMs focus on the conditional distributions that imply that joint. Thus, the negpotential function must be constructed for a LSGM. The form of the H functions in the expansion (4.8) in terms of conditional probability distributions is given by Kaiser and Cressie (2000). Neither ERGMs nor LSGMs consider all terms in the expansion in (4.8).

Corollary 4.3.2 states that the specific ERGM with density and two-star parameters induces an uncentered parameterization. However, an examination of the proof will show that, although the conditional distributions and corollary were proven for this particular ERGM specification, a similar proof would result for any ERGM which contained a density parameter and including a density parameter into an ERGM has been stated to be analogous to including an intercept in a regression model.

4.4.1 Centering of Third Summation in LSGM

The Besag (1974) parameterization of conditional exponential family distributions (including the auto-regression model (4.2)) was developed for spatial data under an assumption of pairwise-only dependence (cliques of size at most two in (4.8)), and it was this MRF model form that was extended to a centered version by Kaiser et al. (2012). Lee et al. (2001) extended Besag's original work to give a parameterization suitable for inclusion of cliques of any size, but did not consider how their parameterization might be centered. In this section we will develop a centered parameterization for conditional exponential family distributions when the pairwise-only dependence assumption is dropped and dependence between triples of random variables is explicitly allowed, which can then be applied to formulating LSGMs with higher-order dependence terms using conditional distributions (4.1) for edge occurrence that go beyond centered parameterizations as in (4.4).

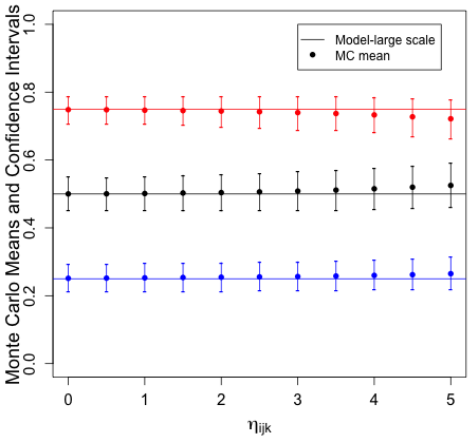
The expanded centering will be explained and first developed for a traditional MRF setting in which random variables have locations indexed by nodes of a regular lattice and edges denote the neighborhood structure. For illustration in the following, we will use a 20×20 regular lattice, so that the $m = 400$ random variables are assigned at regularly-spaced grid locations. We will assume an eight-nearest neighborhood structure and will wrap the grid on a torus so that all random variables on the lattice have eight neighbors.

Let $\mathcal{C}_i^3 = \{\mathbf{s}_j; \mathbf{s}_j, \mathbf{s}_i \text{ and } \mathbf{s}_k \text{ are neighbors for } j \neq i \text{ and some } k \notin \{i, j\}\}$ be all cliques of size three corresponding to the random variable $Y(\mathbf{s}_i)$. A centered natural parameter function in (4.1) which incorporates dependence between triples of random variables is

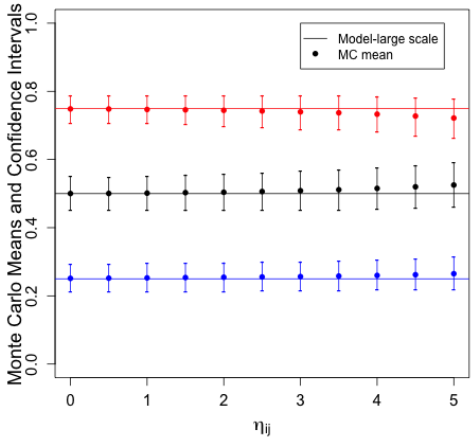
$$A_i(N_i) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i} \eta_{ij}[y(\mathbf{s}_j) - \kappa_j] + \sum\sum_{\mathbf{s}_j, \mathbf{s}_k \in \mathcal{C}_i^3} \eta_{ijk}[y(\mathbf{s}_j)y(\mathbf{s}_k) - \kappa_j\kappa_k] \quad i = 1, \ldots, m \tag{4.9}$$

where the parameter η_{ij} represents the amount of dependence between $Y(\mathbf{s}_i)$ and each of its neighbors individually, and the parameter η_{ijk} represents the amount of dependence between $Y(\mathbf{s}_i)$ and the other variables in cliques of size three to which it belongs. Simplifying assumptions are necessary to decrease the number of parameters, such as a single $\kappa = \kappa_i$, or $\eta_{ij} \equiv \eta_2$ for all $\mathbf{s}_j \in N_i$ and $\eta_{ijk} \equiv \eta_3$ for all pairs $\mathbf{s}_j, \mathbf{s}_k \in \mathcal{C}_i^3$.

Two simulation studies were performed on a 20×20 regular lattice to demonstrate the near equality between κ_i and the marginal mean of $Y(\mathbf{s}_i)$ across various values of the two dependence parameters, η_{ij} and η_{ijk}. For simplicity, each network was simulated with a fixed large-scale parameter $\kappa_i = \kappa$, a fixed two-way dependence parameter $\eta_{ij} = \eta_2$, and a fixed three-way dependence parameter $\eta_{ijk} = \eta_3$. These simulations have the same intent as those of Caragea and Kaiser (2009) for binary models assuming pariwise-only dependence. For each combination of parameter values, 10,000 lattices were simulated using a Gibbs sampler with a burn-in of 10,000 and a thinning value of 25. Simulation summaries are displayed in Figure 4.1. The plot on the left displays the results when simulating from a binary MRF model (4.9) on a 20×20 regular lattice when the value of the two-way dependence is fixed at $\eta_2 = 2$ and for eleven values of η_3. This was repeated for three values of a fixed marginal mean, $\kappa = \{0.25, 0.5, 0.75\}$. The horizontal lines in the plot represent the simulated κ parameter as



(a) For $\eta_{ij} = 2$ and $\kappa = 0.25$ (blue), $\kappa = 0.5$ (black) and $\kappa = 0.75$ (red) and based on 10,000 simulated networks for each of 11 values of η_{ijk}.



(b) For $\eta_{ijk} = 2$ and $\kappa = 0.25$ (blue), $\kappa = 0.5$ (black) and $\kappa = 0.75$ (red) and based on 10,000 simulated networks each of 11 values of η_{ij}.

Figure 4.1 Simulation study to show the effect of the centering of the third-order term on a 20×20 lattice. Points represent the average proportion of realized edges (as an approximation of marginal expectation $E[Y(\mathbf{s}_i)]$) with 90% confidence intervals.

$\kappa = 0.25$ in blue, $\kappa = 0.5$ in black, and $\kappa = 0.75$ in red. The points represent the average proportion of simulated random variables realized with 90% confidence intervals. As can be seen, as the value of η_3 increases from 0 to 5, the simulated marginal mean remains nearly equal to κ. The plot on the left displays the results of a simulation study with the value of $\eta_3 = 2$ and the two-way dependence parameter, η_2, increasing from 0 to 5. Again, due to the centered parameterization, the marginal means of the simulated values remain nearly equal to the large-scale model parameter κ.

As mentioned previously in Section 4.2, one way in which Caragea and Kaiser (2009) demonstrate the advantages of a centered parameterization for binary models is to compare the log odds ratio of both centered and uncentered parameterizations relative to the independence model. An independence model results from $\eta_{ij} = \eta_{ijk} = 0$ in (4.9), as in the pairwise-only dependence scenario. Thus, the log odds ratio for the expanded centered model from (4.9) is

$$\log\left[\frac{c_i/(1-c_i)}{p_i/(1-p_i)}\right] = \sum_{\mathbf{s}_j \in N_i} \eta_{ij}[y(\mathbf{s}_j) - \kappa_j] + \sum\sum_{\mathbf{s}_j, \mathbf{s}_k \in \mathcal{C}_i^3} \eta_{ijk}[y(\mathbf{s}_j)y(\mathbf{s}_k) - \kappa_j\kappa_k]$$

demonstrating that the odds that $Y(\mathbf{s}_i) = 1$ increases if the number of positive neighbors is more than expected under the independence model, assuming that $\eta_{ijk} > 0$, and decreases if the number of neighbors is less than expected. This illustrates the effect of the small-scale model structure on the conditional expectation. When many neighbors of an edge are not realized, the conditional expectation $c_i = E[Y(\mathbf{s}_i)|\mathbf{y}(N_i)] = \Pr(Y(\mathbf{s}_i) = 1|\mathbf{y}(N_i))$ of edge occurrence is often less than the marginal expectation, $E[Y(\mathbf{s}_i)]$ (where $E[Y(\mathbf{s}_i)] \approx \kappa_i$ with parameter centering and reasonable dependence parameters as in Figure 4.1) as will be demonstrated in the two examples of Section 4.5. As the number of positive neighbors increases, the conditional expectation of edge occurrence will also typically increase.

4.5 Example

4.5.1 Inclusion of Attributes

Our first data example demonstrates the inclusion of auxiliary information in both the large and small scale model structures in a LSGM (corresponding to a binary MRF model on graph edges). Data used for this example were obtained from the Buell-Small succession Study (Pickett, 1982), which tracked changes in the vegetation of abandoned agricultural fields beginning in 1958 through the presence or absence of 476 species. Annual visual examination of permanent plots within the fields, located in Somerset County, New Jersey, was used to determine the presence or absence of each species. For more information on the study, see Buell et al. (1971), Caragea and Kaiser (2009), and Pickett (1982).

Species under consideration here are limited to red sorrel and Japanese honeysuckle. Both species are expected to compete for similar resources, as red sorrel is a perennial weed, and Japanese honeysuckle is an invasive, exotic vine. The fields of interest consist of 93, 2.0×0.5 meter, rectangular plots located mostly on a grid. Nodes in the network are defined by the 93 plots, with a separate node for each year and species combination. Years of the Buell-Small succession study where neither red sorrel nor Japanese honeysuckle were completely absent are from 1969 to 1978, and we have limited our attention to those years. Edges form between nodes representing the same physical plot and species for two successive years. The edge is realized if

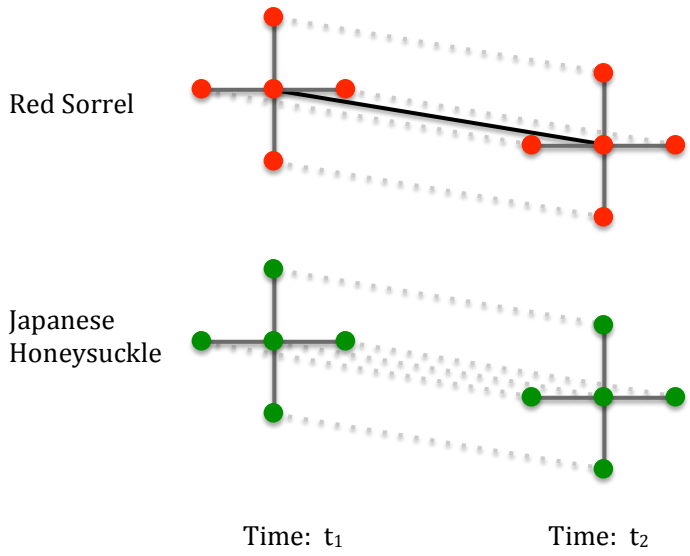


Figure 4.2 Explanation of neighborhood definition used in the Buell-Small succession study example. Each sets of five plots represent the same spatial locations. The black, solid line represents the focal edge, and the gray, dashed lines are its neighbors.

the species is observed at that location in both years. There are 1674 potential edges between the 93 spatial locations for 9 successive year spans and both species.

Neighborhoods are motivated by the spatial four-nearest neighbor definition. Most edges will have nine neighbors defined as types of edges: one for the opposite species at the same location, four that connect the same species from the four nearest spatial plots, and four that connect the opposite species at the four nearest spatial plots. Those nodes which are located on the boundary of the fields will have less than nine neighbors. An illustration of the neighborhood of a non-boundary edge is displayed in Figure 4.2. In this figure, each set of five points represent the same five spatial plots which are nodes of this graph. The nodes on the left represent year, t_1, where those on the right are for the subsequent year, t_2. The focal edge is the black, solid line connecting the red sorrel nodes in the center. Its neighbors are the four edges connecting red sorrel at the spatially-bordering plots and the five edges connecting Japanese Honeysuckle at the same location as that used to define the focal edge in this example, and the spatially-bordering plots.

Species information will be incorporated into the dependence and large-scale model structure, through two large-scale parameters (termed as κ_{JH} and κ_{RS}) and pairwise dependence parameters (termed as η_S and η_O). The parameters κ_{JH} and κ_{RS} allow for the marginal expectation to vary whether the potential edge connects two Japanese honeysuckle (JH) or two red sorrel (RS) nodes. For the two pairwise-dependence parameters η_S and η_O, neighbors are classified as one of two types, either corresponding to the same or opposite species. In Figure 4.2, the four neighbors between the red nodes are of one type, and the five edges between the green nodes are of the other type. Following the recommendation of Casleton et al. (2014b), both pairwise dependence parameters will be adjusted to account for the unequal neighborhood sizes. The natural parameter function for the potential edge $Y(\mathbf{s}_i)$ in a LSGM (a binary MRF on edges) with a centered parameterization is

$$A_i\{\mathbf{y}(N_i)\} = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i^S} \frac{\eta_S}{|N_i^S| + |N_j^S|}[y(\mathbf{s}_j) - \kappa_i] + \sum_{\mathbf{s}_k \in N_i^O} \frac{\eta_O}{|N_i^O| + |N_k^O|}[y(\mathbf{s}_k) - \kappa_{i^c}]$$

where $\kappa_i = \kappa_{JH}$ or κ_{RS} represents the large-scale parameter which corresponds to the species of $Y(\mathbf{s}_i)$, $|N_i^S|$ represents the number of neighbors of $Y(\mathbf{s}_i)$ that are of the same species, $|N_i^O|$ is the number of neighbors of the other species, and $\kappa_{i^c} = \{\kappa_{RS}, \kappa_{JH}\} \cap \{\kappa_i\}^c$ is the large-scale parameter for the species opposite of that of $Y(\mathbf{s}_i)$. It should be noted that this particular neighborhood definition does not lead to any cliques of size three, and thus, by the Hammersly-Clifford theorem, no η_3 parameter is included in the model.

Parameter estimates were obtained by maximizing the log pseudo-likelihood (PL) of Besag (1975). The log PL for the LSGM is

$$\log \mathrm{PL} = \sum_i \{y(\mathbf{s}_i) \log[p_i(N_i)] + (1 - y(\mathbf{s}_i)) \log[1 - p_i(N_i)]\}$$

where $p_i(N_i)$ represents the conditional expectation for $Y(\mathbf{s}_i)|\mathbf{y}(N_i)$,

$$p_i(N_i) = \frac{\exp(A_i\{\mathbf{y}(N_i)\})}{1 + \exp(A_i\{\mathbf{y}(N_i)\})}. \tag{4.10}$$

The maximization of the PL function for the Buell-Small succession network results in the estimates of Table 4.1. The estimate of $\hat{\kappa}_{JH} = 0.13$ for Japanese honeysuckle is less than that of $\hat{\kappa}_{RS} = 0.38$ for red sorrel representing the higher prevalence of the weed compared to the

Table 4.1 Parameter estimates and 90% percentile parametric bootstrap confidence intervals for the LSGM fit to the Buell-Small succession network.

	Estimate	90% Confidence Interval
κ_{JH}	0.13	(0.06, 0.39)
κ_{RS}	0.38	(0.13, 0.63)
η_S	6.43	(5.39, 7.41)
η_O	-4.17	(-5.47, -3.20)

exotic vine over the area of interest, although interval estimates are wide with considerable overlap. As expected, the dependence parameter for pairs of neighbors of differing species is negative, while the pairwise dependence for same species is positive, and the interval estimates of these two parameters are well separated.

Although estimates from the PL function for a MRF are generally consistent and asymptotically normal (Guyon, 1995), standard errors of the estimates can be difficult to compute. Thus, confidence intervals were estimated using a percentile parametric bootstrap. One-thousand networks were simulated from the estimates in Table 4.1 through a Gibbs sampler with a burn-in of 10,000 with an equal number of simulations between each network retained. Parameter estimates were obtained for each simulated network through maximum PL resulting in 90% confidence intervals from the 5th and 95th percentiles of the empirical distributions of these estimates. These intervals are also shown in Table 4.1.

To aid in interpretation of the fitted model, Table 4.2 displays conditional expectations, (4.10), for both species for the 30 possible configurations of neighboring values, which were computed using the estimates of Table 4.1. Because of the centered parameterization, the marginal expectation for each species is taken as the estimate of the corresponding large-scale parameter κ, so 0.13 for Japanese honeysuckle and 0.38 for red sorrel. Conditional expectations depend on the number of positive neighbors and whether those neighbors are of the same species or the opposite species. Those values displayed in red in Table 4.2 are the conditional expectations less than the corresponding marginal expectation. These values typically occur when the number of opposite species present is greater than the number of same species. Conditional expectations for red sorrel are more likely than Japanese honeysuckle to be less than

Table 4.2 Conditional expectations for both Japanese honeysuckle and red sorrel edges based on the number of positive neighbors. Values in red are less than the corresponding marginal expectations.

Japanese Honeysuckle							Red Sorrel						
# positive	# positive opposite species						# positive	# positive opposite species					
same species	0	1	2	3	4	5	same species	0	1	2	3	4	5
0	0.18	0.13	0.09	0.06	0.04	0.03	0	0.19	0.14	0.09	0.06	0.04	0.03
1	0.33	0.25	0.18	0.12	0.09	0.06	1	0.35	0.26	0.19	0.13	0.09	0.06
2	0.53	0.42	0.33	0.24	0.17	0.12	2	0.54	0.44	0.34	0.25	0.18	0.13
3	0.71	0.62	0.52	0.42	0.32	0.24	3	0.73	0.64	0.54	0.43	0.33	0.25
4	0.85	0.79	0.71	0.61	0.51	0.41	4	0.86	0.80	0.72	0.63	0.53	0.43

the respective marginal expectation, which can be attributed to the larger marginal expectation estimated for red sorrel.

To assess the model fit to the data, we will examine how well simulated networks are able to recreate a specific feature of the realized network. The network feature of interest is the proportion of neighbors of the same species which are realized. This feature was chosen because it is not the neighborhood definition used to build the model, even though the feature is used as part of the definition. The proportion realized is considered, rather than the number of neighbors, due to the variable neighborhood sizes resulting from boundary effects. In the Buell-Small succession network, the average proportion of these type of neighbors realized is 0.297. For each of the 1000 simulations obtained from the fitted model, the average proportion of neighbors of the same species for each potential edge was computed. Figure 4.3 displays the averages from the simulations. The red, dashed vertical line represents the mean from the realized network which appears near the center of the distribution. In fact, if we use the simulated distribution as a reference distribution to test the hypothesis of model fit, using the mean from the realized network as the test statistic, the resulting p-value is 0.648. Thus, there is strong evidence the LSGM is able to capture this network feature.

4.5.2 Inclusion of Higher-Order Terms

To demonstrate the inclusion of cliques of size three into a LSGM, we will use networks constructed from American football games between NCAA Division I universities from two

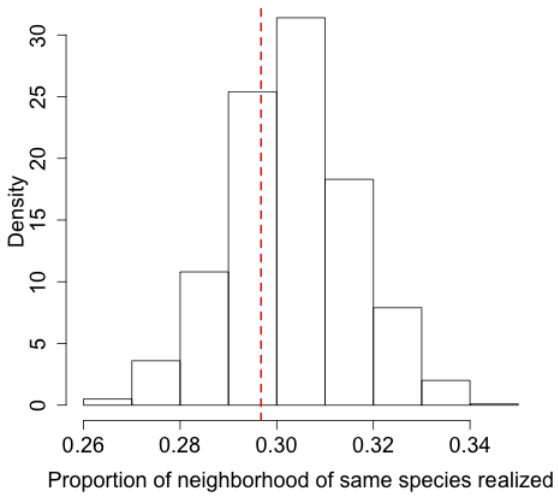


Figure 4.3 Proportion of neighbors of the same species which are realized in 1000 simulations from the model fit in Table 4.1.

seasons. The two seasons of interest are 2000 and 2013, where the nodes represent the 115 schools with a Division I college football team in 2000. The locations of the nodes of both networks are displayed in Figure 4.4 where the position of the University of Hawai'i has been modified to fit onto the map. An edge is realized between two nodes in the network if the schools competed against each other in that particular season.

Most college football programs are a member of an athletic conference, a group of teams who predominately compete against each other (with the exception of schools classified as Independent (Indep) who do not belong to any conference). The network of games from the 2000 season was compiled by Girvan and Newman (2002) and has been used as a test network to evaluate community detection techniques by using the network to uncover conference structure (Guo et al., 2013). This is not the purpose of this example, so the conference information is considered known.

One of the main functions of collegiate football athletic conferences is to negotiate television contracts. Recently, conferences have looked to expand into new media markets or to include schools with a large fan base in order to increase revenue. This results in larger conferences, and the shift between 2000 and 2013 is displayed in Figure 4.4. The same 115 schools from the

original 2000 season will be considered, even though other schools have since acquired Division I status. Games played between the 115 schools resulted in 613 realized edges for the 2000 season and 598 in 2013.

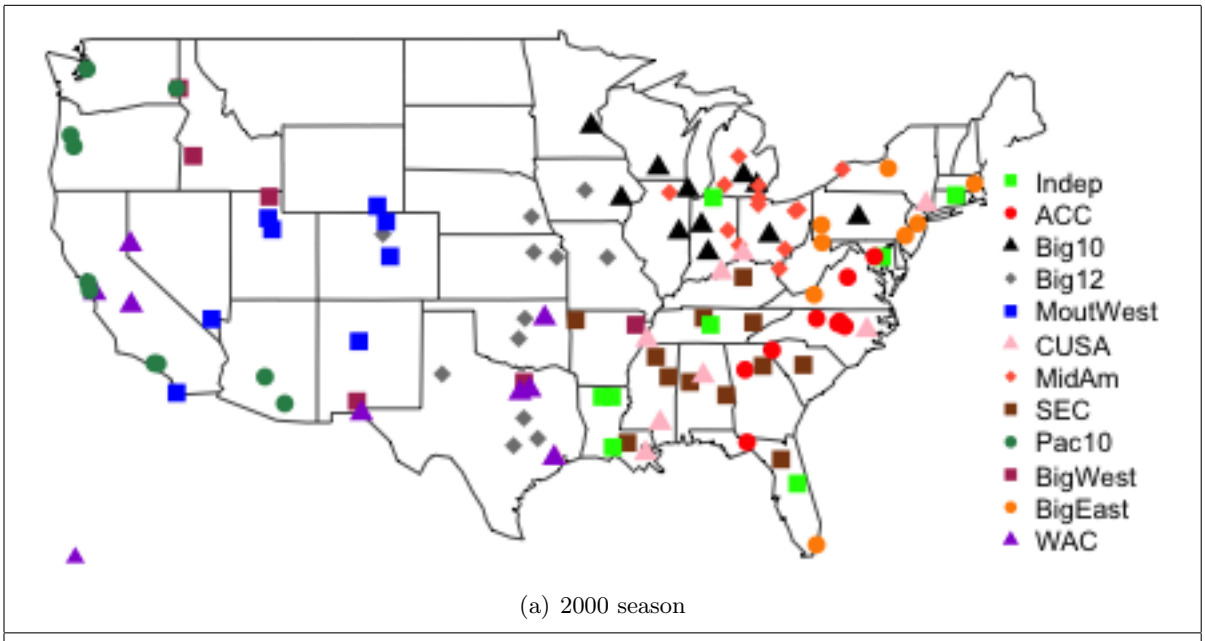
Although a school's football schedule predominantly contains other schools within the same conference, a school will play teams from other conferences, particularly at the beginning of a season. Thus, we will allow an edge to form between any two possible nodes leading to $\binom{115}{2} = 6555$ possible edges in both networks. Neighborhoods will be defined only for edges which connect schools in the same conference. Edges that result from out-of-conference games or games where one team is Independent will not have neighbors and thus will be modeled as occurring independently of any other possible edges. Two edges will be considered neighbors if they are incident, i.e., corresponding to different games of a single school, and correspond to games among schools within the same conference.

The model for both 2000 and 2013 football networks will contain two large-scale parameters: one κ_{out} for out-of-conference games, and one κ_{in} for edges which form between in-conference opponents. This is intuitive as the proportion of possible in-conference games realized is much higher than the proportion of possible out-of-conference games that are played. The natural parameter function for an in-conference edge is given as

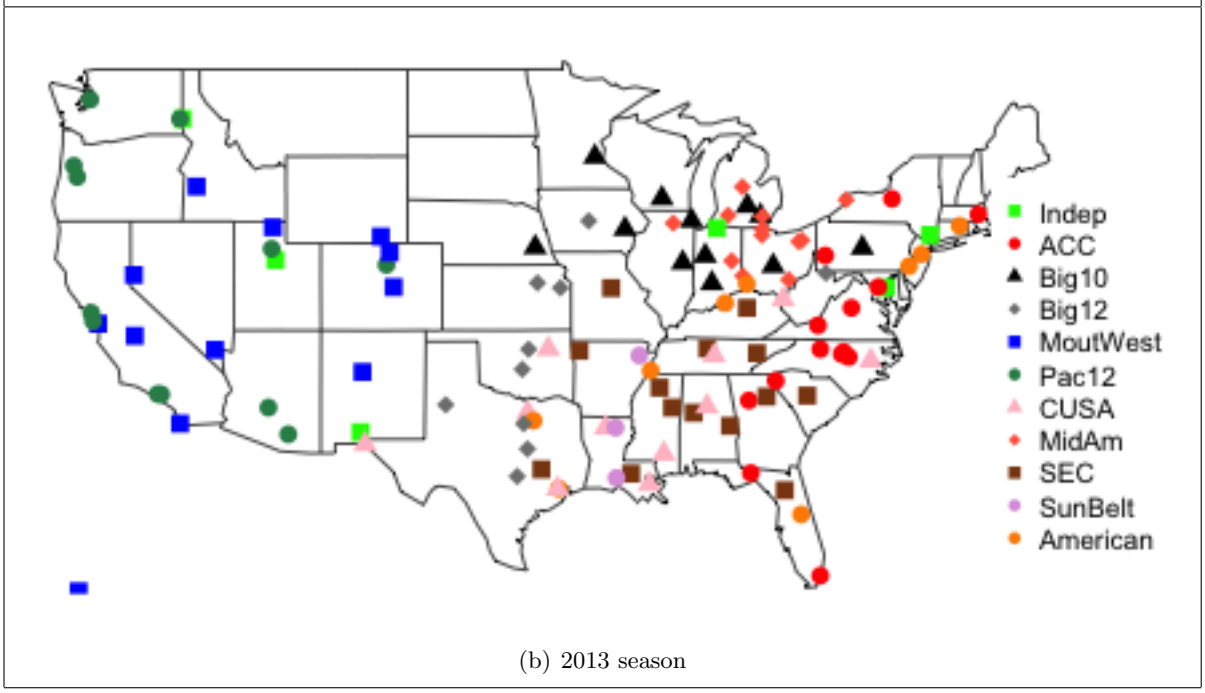
$$A_i\{\mathbf{y}(N_i)\} = \log\left(\frac{\kappa_{\text{in}}}{1-\kappa_{\text{in}}}\right) + \sum_{\mathbf{s}_j,\mathbf{s}_k \in \mathcal{C}_i^3} \left[\frac{\eta_3}{|\mathcal{C}_i^3| + |\mathcal{C}_j^3| + |\mathcal{C}_k^3|}\right] (y(\mathbf{s}_j)y(\mathbf{s}_k) - \kappa_{\text{in}}^2)$$

where $|\mathcal{C}_\ell^3|$ represents the number of cliques of size three to which the edge $y(\boldsymbol{s}_\ell)$ belongs. A common η_3 parameter, which represents the amount of dependence from cliques of size three, is adjusted for potentially varying sizes in the number of cliques of size three to which an edge belongs. Dividing the parameter η_3 by this sum allows the second term in $A_i\{\mathbf{y}(N_i)\}$ to have a uniform effect on the natural parameter function while preserving the symmetry of the parameters, a requirement for the identification of a joint distribution (Casleton et al., 2014b).

Parameter estimates for the 2000 and 2013 seasons are displayed in Table 4.5.2. Again, point estimates were obtained through maximum PL and confidence intervals through a percentile parametric bootstrap of 1000 simulated networks with burn-in and thinning values of 10,000 each. The estimated marginal expectation for out-of-conference games is similar between the



(a) 2000 season



(b) 2013 season

Figure 4.4 Nodes of the college football network and their classification based on conference (with a slight geographic adjustment to the University of Hawai'i).

Table 4.3 Parameter estimates and 90% percentile parametric bootstrap confidence intervals for the football networks.

	2000 season	2013 season
κ_{out}	0.034 (0.031, 0.038)	0.030 (0.027, 0.034)
κ_{in}	0.831 (0.794, 0.873)	0.716 (0.685, 0.747)
η_3	3.97 (-1.75, 7.79)	-0.181 (-5.88, 3.15)

two seasons, which is expected given that the proportion of possible out-of-conference games played stayed constant between the two years. For in-conference games, the marginal expectation significantly decreased from the 2000 to 2013 season. This can be attributed to the increasing size of conferences and thus a smaller proportion of possible in-conference games being played.

The point estimate of the parameter η_3 in 2013 did decrease from its value in 2000, and the estimate for the 2013 season is, for all practical purposes, zero. This implies there is little dependence between triples of random variables in the 2013 season. To further demonstrate this difference, Figure 4.5 plots the conditional and marginal expectations for both fitted models using the point estimates in Table 4.5.2. Conditional expectations are computed for a random variable included in 100 cliques of size three, the most common size for both datasets. Marginal expectations for an in-conference game are given by the estimated κ_{in} values, while conditional expectations depend on the number of cliques of size three that are realized. The conditional expectations for the 2000 season range from 0.66 when none of the cliques of size three are realized to 0.88 when all cliques of size three are present. After about 70% of the cliques of size three are present, the conditional expectation exceeds the marginal expectation. For the 2013 season, the conditional and marginal expectations are very similar regardless of the number of cliques of size three realized, indicating the small amount of dependence. This decrease in dependence can also be attributed to larger conferences as fewer cliques of size three are realized due to the inability of teams within a conference to play everyone else.

A larger η_3 value indicates a stronger amount of dependence between the cliques of size three for the 2000 season than the 2013 season. The feature of the data which this parameter

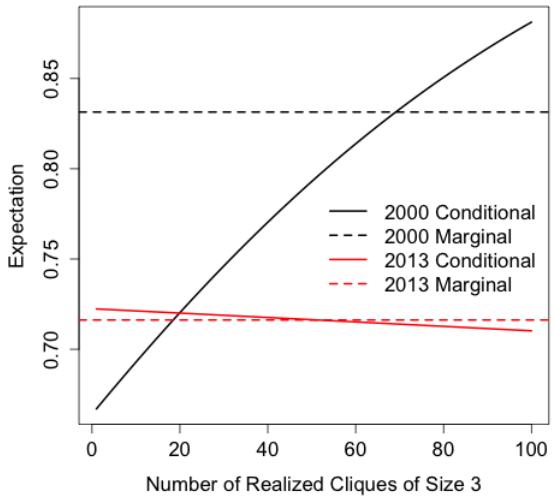


Figure 4.5 Marginal and conditional expectations for the fitted models to the 2000 and 2013 NCAA college football networks.

is attempting to capture is demonstrated in the structure of the cliques of size three. Consider an edge for a marker/game, $\mathbf{s}_i$ and assume the value of this "focal" edge is $y(\mathbf{s}_i) = v$ where $v \in \{0, 1\}$. Each of the cliques of size three to which this edge belongs contains two edges other than that of $\mathbf{s}_i$. To characterize the structure of the cliques of size three, we will examine the number of edges in each pair which assume the value v. This results in three proportions: the proportion of cliques of size three where neither edge assumes the value v, the proportion of cliques of size three where exactly one of the two edges takes the value v, and lastly, the proportion of cliques of size three where all three edges have the same value.

The proportions which characterize the cliques of size three for the 2000 and 2013 football networks are displayed in Table 4.5.2. These values can be interpreted as probabilities for the number of edges in the clique of size three that will take the same value as a randomly chosen focal edge. If an edge in the 2000 season football dataset is present (absent), it is most likely that the other two edges in the clique of size three are also present (absent). This is less true in the 2013 football season where the probabilities are similar for all three possibilities and highest for only one other edge in the clique assuming the same value.

Table 4.4 Characterization of the cliques of size three for the 2000 and 2013 football networks as the proportion of cliques of size three for which none, one, or both other edges assume the same value as the focal edge and p-values for the distributions in Figure 4.6.

	Averages			p-values	
# Same as edge of interest	2000	2013		2000	2013
none	0.14	0.21		0.47	0.37
one	0.28	0.41		0.47	0.36
both	0.58	0.38		0.52	0.63

Model assessment can be accomplished by examining how well the simulations from the fitted models are able to recreate the structure of the cliques of size three in the realized networks. For both fitted models, 1000 simulated networks are obtained. The proportion of cliques of size three with none, one, or two values the same as a focal edge are computed for each simulation as in Table 4.5.2. Proportions from the simulations are displayed as boxplots in Figure 4.6, and the red points represent the proportions from the actual networks. Because the realized values fall within the middle of the distributions of simulated values, the simulated networks are able to adequately recreate this feature of the data. If we again use the simulated distributions as reference distributions and the observed values as the test statistics for a test of model fit, the resulting p-values are shown on the right of Table 4.5.2. All p-values are large which indicates there is no issue with lack of fit with respect to this model feature.

4.6 Conclusions

Two possible extensions to local structure graph models (LSGMs) have been presented and their use illustrated with two example networks. As a first extension, auxiliary information, specifically covariate information on the nodes, was incorporated into the large scale structure through different large-scale model parameters, (e.g., for various species or inter- and intra-conference games as in Sections 4.5.1 and 4.5.2, respectively), into the small scale structure through different dependence parameters (e.g., based on neighbors which connect the same

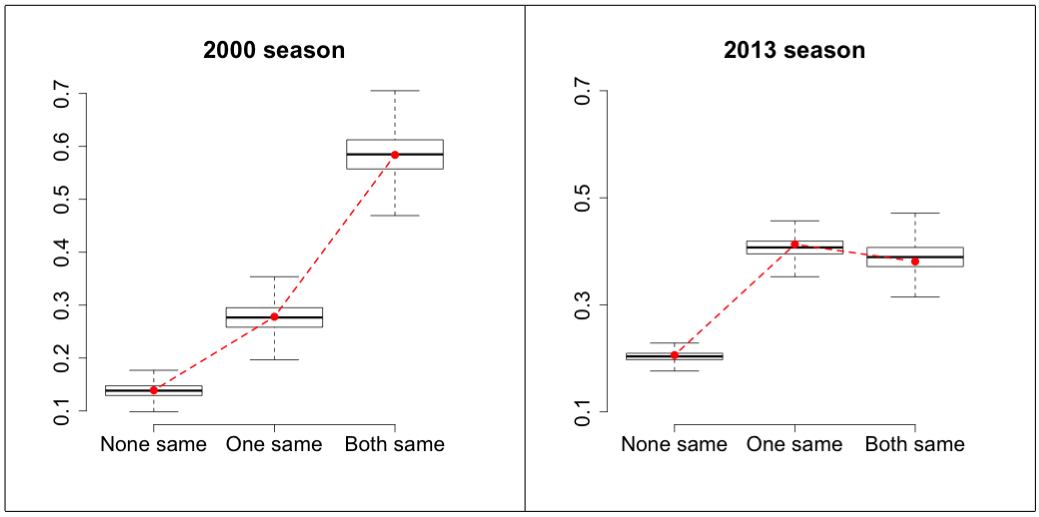


Figure 4.6 Model assessment of the fits to the 2000 and 2013 football datasets. Boxplots represent proportion of cliques of size three with the corresponding number of edges having the same value as the edge of interest from 1000 simulated networks. Red points represent the proportions from the realized networks.

or a competing species), or into the dependence (e.g., different neighborhoods for inter- and intra-conference games).

The second extension relaxes the pairwise-only dependence assumption in the binary conditional specification of LSGMs to allow for higher order dependence and account for dependence between triples of random variables. This requires an extension of the centered parameterization presented by Caragea and Kaiser (2009), Hughes et al. (2011), and Kaiser et al. (2012) for Markov Random Field (MRF) models. An advantage of the centered parameterization over the original version is a more uniform interpretation of parameters across sensible levels of statistical dependence. This advantage in interpretability is maintained when centering the higher order dependence terms is used in the manner presented here. The centering extension is applicable to the more general MRF models, of which LSGMs are a special case. In addition, cases of exponential random graph models (ERGMs) were shown to induce conditionals distributions with uncentered parameterizations, which indicates that these models can exhibit confounded parameters that are difficult to interpret.

Although not explicitly demonstrated in this work, another advantage in the increased interpretability of parameters is the ability to better recognize the region of a parameter space which can lead to model degeneracy. Currently, the suggestion to identifying model degeneracy in LSGMs is to simulate from the fitted model and verify the simulated proportion of edges is reasonable compared to those from the realized network. The recognition of model degeneracy through a more rigorous approach is the subject of future work.

4.7 Appendix

4.7.1 Proof of Proposition 4.3.1

Proof. Let $N_i^- = \{\mathbf{s}_i, N_i\}^c$ be the edges which are not neighbors of $y(\mathbf{s}_i)$ or the edge $y(\mathbf{s}_i)$ itself. The log of the joint distribution for the ERGM with a density and two-star parameter can be written as proportional to

$$\begin{aligned}\log[\Pr(\mathbf{y})] \propto \rho &\left[y(\mathbf{s}_i) + \sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j) + \sum_{\mathbf{s}_k \in N_i^-} y(\mathbf{s}_k) \right] \\ &+ \frac{\sigma}{2} \left[y(\mathbf{s}_i) S_i + \sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j) S_j + \sum_{\mathbf{s}_k \in N_i^-} y(\mathbf{s}_k) S_k \right]\end{aligned}$$

where the sums have been separated based on if they contain neighbors of $y(\mathbf{s}_i)$.

The form of the marginal of $\mathbf{y}(N_i)$ and $y(\mathbf{s}_i)$ can be determined by summing over all possible combinations of the possible values, 0 and 1, of the edges in N_i^-.

$$\begin{aligned}\log[\Pr(\mathbf{y}(N_i), y(\mathbf{s}_i))] \propto \rho &\left[y(\mathbf{s}_i) + \sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j) + c_1 \right] \\ &+ \sigma \left[y(\mathbf{s}_i) S_i + \sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j) S_j^- + \sum_{\substack{\mathbf{s}_k \in N_i \\ \mathbf{s}_k \in \{N_g : \mathbf{s}_g \in N_i^-\}}} c_{2k} y(\mathbf{s}_k) + c_3 \right] \\ &= y(\mathbf{s}_i) D_i + D_j^{-i}\end{aligned}$$

where c_1, c_{2k}, and c_3 are constants, and

$$S_i = \sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j)$$

$$D_i \equiv A_i\{\mathbf{y}(N_i)\} = \rho + \sigma S_i$$

$$S_j^- = \sum_{\substack{\mathbf{s}_k \in N_j \\ \mathbf{s}_k \neq \mathbf{s}_i \\ \mathbf{s}_k \notin N_i^-}} y(\mathbf{s}_k)$$

$$D_j^{-i} = \rho \left[\sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j) + c_1 \right] + \sigma \left[\sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j) S_j^{-i} + \sum_{\substack{\mathbf{s}_k \in N_i \\ \mathbf{s}_k \in \{N_g : \mathbf{s}_g \in N_i^-\}}} c_{2k} y(\mathbf{s}_k) + c_3 \right]$$

Note that in each of these terms, the product(s) that include the random variable $y(\mathbf{s}_i)$ has been removed. It also should be noted that leading parameter of the second term that counts the number of two-stars, the product $y(\mathbf{s}_\ell)S_\ell$ will appear twice for all possible ℓ.

Also, the marginal distribution of only $\mathbf{y}(N_i)$ is

$$\Pr(\mathbf{y}(N_i)) \propto \exp[D_j^{-i}][1 + \exp[D_i]]$$

Therefore,

$$\begin{aligned} Pr(y(\mathbf{s}_i)|\mathbf{y}(N_i)) &= \frac{Pr(\mathbf{y}(N_i), y(\mathbf{s}_i))}{Pr(\mathbf{y}(N_i))} \\ &= \frac{\exp[D_j^{-i}] \exp[y(\mathbf{s}_i) D_i]}{\exp[D_j^{-i}](1 + \exp(D_i))} \\ &= \frac{\exp[y(\mathbf{s}_i) D_i]}{1 + \exp[D_i]} \end{aligned}$$

□

4.7.2 Proof of Corollary 4.3.2

Proof. To prove the corollary, we will compute the log odds ratio relative to the independence model and compare the result to the uncentered in (4.3) and the centered in (4.5). The independence model will correspond to a model with only the density parameter, ρ. Again, let $c_i = Pr(y(\mathbf{s}_i)|\mathbf{y}(N_i)) = Pr(y(\mathbf{s}_i) = 1)$ under the independence model and $p_i = Pr(y(\mathbf{s}_i) =$

$1|\mathbf{y}(N_i))$ under theERGM with a density and two-star parameter. Then,

$$\log(c_i/(1-c_i)) = \rho$$

For the ERGM with density and two-star parameters and with $D_i \equiv A_i\{\mathbf{y}(N_i)\}$ as above,

$$\begin{aligned} Pr(y(\mathbf{s}_i) = 1|\mathbf{y}(N_i)) &= \frac{\exp[D_i]}{1+\exp[D_i]} \\ Pr(y(\mathbf{s}_i) = 0|\mathbf{y}(N_i)) &= \frac{1}{1+\exp[D_i]} \\ \log(p_i/(1-p_i)) &= D_i = \rho + \sigma S_i \end{aligned}$$

Thus,

$$\begin{aligned} \log\left[\frac{c_i/(1-c_i)}{p_i/(1-p_i)}\right] &= -\sigma S_i \\ &= -\sigma \sum_{\mathbf{s}_j \in N_i} y(\mathbf{s}_j) \end{aligned}$$

Thus, this corresponds to the uncentered parameterization in (4.3) with constant pairwise spatial dependence parameters, $-\sigma = \eta_{ij},\ \forall\ i, j$.

□

CHAPTER 5. DATA STRUCTURES REPRESENTED BY A RANDOM GRAPH MODEL: WHEN IS TRANSITIVITY NEEDED?

5.1 Introduction

Transitivity, or the existence of triangles within a network, has drawn a lot of attention in network analysis. Within the fields of computer science and statistical physics, the focus of many analyses is on creating algorithmic, graph generators which emulate an observed network as closely as possible. Clustering, the more common name for transitivity in this field (Watts and Strogatz, 1998), was identified as one of the three features that graph generators should specifically aim to recreate (Lancichinetti et al., 2008). With respect to social networks, transitivity has a natural interpretation and thus makes it a desirable feature to describe with a parameter in a model. The interpretation is that if two individuals share a common friend, they are more likely to also be friends. In fact, Newman et al. (2002) claim that the probability that two people are friends is several orders of magnitude greater if they have a common connection over those that do not. Another argument has been made that this feature alone separates realized social networks from those networks that have been generated at random (Snijders et al., 2006). When the inclusion of transitivity was suspected to lead to model degeneracy (Robins et al., 2007), steps were taken to modify the way it was incorporated into the model so that it could still be accounted for (Hunter et al., 2008a; Hunter and Handcock, 2006; Snijders et al., 2006).

With all the attention given to transitivity, more focus has been on how to include it in a model rather than on if it should be included in the model. The focus of this chapter is to explore the question of including transitivity with two common example networks. The first is a friendship network between high-school aged students, commonly referred to as Goodreau's

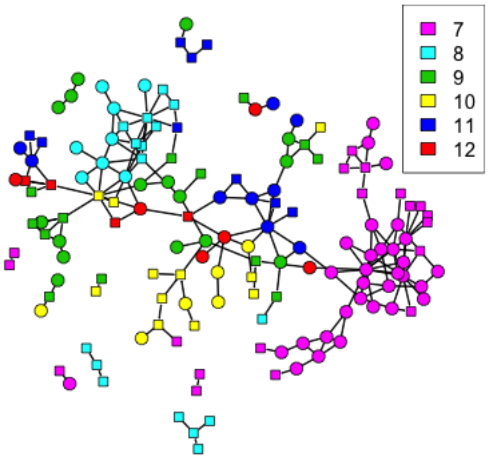
Faux Mesa High friendship network (Handcock et al., 2014), and the second is constructed from the games of one NCAA Division-I college football season (Girvan and Newman, 2002). These two networks were chosen because they represent two ends of the spectrum in terms of realized percentage of transitivity and other network features of interest. The network modeling approach used to answer this question is the local structure graph model (LSGM). The networks will be examined in detail and then various models, with and without the inclusion of transitivity, are fit to both networks and compared.

The rest of the chapter is organized as follows: Section 5.2 describes the example networks and how they have been used in other works as both networks have been examined elsewhere, but in different contexts than that presented here. The network analysis approach, the LSGM, is described in detail in Section 5.3. Potential and realized topological features of both the high school and football networks are detailed in Section 5.4. Three LSGMs are fit and described for both networks in Section 5.5, and Section 5.6 discusses some implications of the result of the models fit in the previous section.

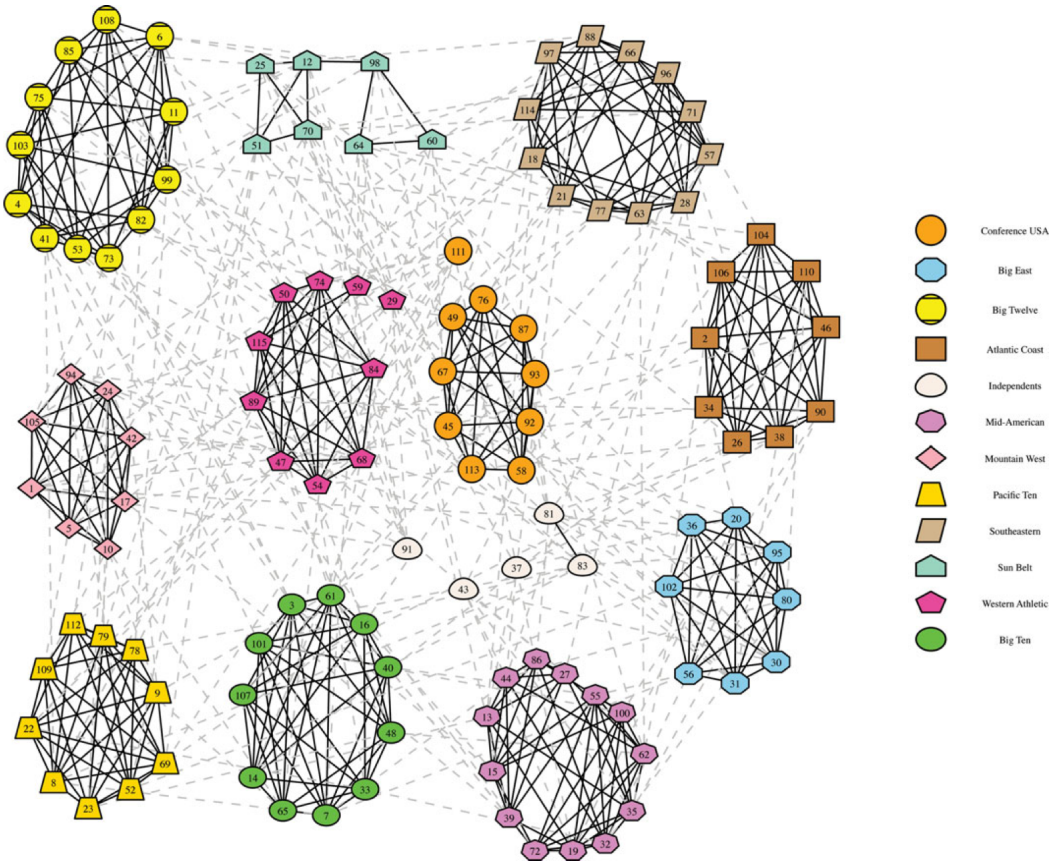
5.2 Example Networks

The first dataset is known as Goodreau's Faux Mesa High friendship network (Handcock et al., 2014). Nodes represent students in grades 7–12 from one high school and corresponding middle school in rural, western United States. Undirected edges form between two nodes if both students identified the other as a friend. This is known as a mutualized friendship network and is a common conceptualization of friendship in social network analysis (Hunter et al., 2008a). The network object is distributed in the `ergm` (Handcock et al., 2014) package for `R` (R Core Team, 2013). A visualization of this network as it appears in Hunter et al. (2008b) obtained from the plotting function of the `ergm` package is shown in Figure 5.1.

The network is based on School 10 (Hunter et al., 2008a) from the first wave of the Add Health longitudinal study. School 10 was chosen by Hunter et al. (2008a) because its analysis was found to be similar to the results obtained from a simultaneous analysis of 59 schools. Add Health is an extensive study with multiple waves of data collection and has been used to analyze adolescent behavior of students from 134 schools through both in-school surveys



(a) Faux Mesa High network. Colors represent the grade and symbol shape represents the sex. Figure adapted from Hunter et al. (2008b).



(b) NCAA Football network. Color and shape of nodes represent the conference. In- and out-of-conference edges are represented differently. Figure taken from Guo et al. (2013).

Figure 5.1 Visualization of the networks of the Faux Mesa High and football network.

and in-home interviews (Resnick et al., 1997). For more information on the survey see http://www.cpc.unc.edu/addhealth/. The friendship network is constructed from the in-school survey where students were given a roster of all other students in the school and asked to list up to five of their closest male and five closest female friends. Nodal attributes of grade, sex, and race (Handcock et al., 2014) are also collected. Race was determined from two questions on race and Hispanic origin with resulting categories: Hispanic, Black, White, Asian, Native American, and Other (Hunter et al., 2008a).

Steps were taken to preserve the confidentiality of the students, so the Faux Mesa High network object does not directly correspond to School 10. First, students who did not take the survey or were not on the school roster were removed and any missing nodal attributes were imputed with random samples from a weighted distribution of known attributes. An ERGM with terms that account for the density, attribute information, and transitivity is fit to the complete data, and the Faux Mesa High network is a single simulation of the fitted ERGM (Goodreau et al., 2008).

Although the Faux Mesa High network is simulated, it has been argued to be a realistic representation of an adolescent friendship network (Handcock et al., 2008) and has been used in various works as an example network. Morris et al. (2008) and Bender-deMoll et al. (2008) use the mutualized friendship network to demonstrate various aspects of ERGMs. A method intended to reduce the impact of a virus attack is tested on the network (Kashirin and Dijkstra, 2013), a graph generation algorithm that uses features on the surface of a hyperplane is demonstrated (Lunga and Kirshner, 2011), and a visualization method for the transmission of a disease (Lofgren, 2012) is illustrated using the Faux Mesa High network.

A saturated graph will be used for the LSGM fit to the Faux Mesa High network. It will allow edges to form only between two nodes which are in the same grade. More than 80% of all realized edges in the network are captured by this saturated graph, but the number of random variables to model is only $m = 4{,}174$ compared to 20,910 when all pairs of nodes are modeled.

The second network of interest is constructed from American football games played between NCAA Division I universities during the 2000 season. Nodes represent the 115 schools with a Division I college football team, and an edges exists between two nodes if the teams competed

against each other that season. Most college football programs are a member of an athletic conference, or a group of teams who predominately compete against each other. An exception are schools classified as Independent who do not belong to any conference. A plot of the nodes, realized edges, and corresponding attribute information are displayed in Figure 5.1, which was obtained from Guo et al. (2013).

The football network was compiled by Girvan and Newman (2002) and has been used as a test network to evaluate community detection techniques by using the network to uncover conference structure (Guo et al., 2013). The goal of the analyses is not to determine this feature, but rather, it will be considered a known attribute, and the information will be used to classify edges. The categories of edges are in-conference for games between two nodes from the same conference and out-of-conference for games between two schools from different conferences or when at least one school is independent.

Although a school's football schedule predominately contains other schools within the same conference, a school will play teams from other conferences, particularly at the beginning of a season. Thus, no saturated graph will be used and edges can form between any two possible nodes, allowing for $\binom{115}{2} = 6555$ possible edges.

5.3 Models

A network is defined by a fixed set of n nodes and m possible edges. Assign to each of the m possible edges a binary random variable $Y(\mathbf{s}_i)$, where the marker $\mathbf{s}_i = \{c_i, r_i\}$ indicates the two nodes, c_i and r_i, that the edge would potentially join. Edge values designate the presence $y(\mathbf{s}_i) = 1$, or absence, $y(\mathbf{s}_i) = 0$, of an edge between the node pair. Covariate information on the nodes can be associated with the marker $\mathbf{s}_i$ and will be designated as a possible vector-valued $\mathbf{x}(\mathbf{s}_i)$.

The random graph model used to analyze the two example networks is the local structure graph model (LSGM) approach. This network analysis technique specifies, for each potential edge $Y(\mathbf{s}_i)$, a conditional distribution, $P(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\{y(\mathbf{s}_j) : j \neq i\}), y(\mathbf{s}_i) = \{0, 1\}$ and a dependent sets of edges, known as neighborhoods, N_i. A Markov dependence assumption

simplifies the conditional distributions so that

$$P(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\{y(\mathbf{s}_j) : j \neq i\}) = P(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i))$$

where $\mathbf{y}(N_i)$ represents the value of the neighbors of $y(\mathbf{s}_i)$. The advantage of conditional specification and an explicit neighborhood definition is control over and interpretation of the local structures in the network. The application of these features to network analysis was introduced in Casleton et al. (2014b) and extended to include higher-order dependence in Casleton et al. (2014a).

LSGMs can be interpreted as an alternate method of specifying another, more common random graph model, exponential random graph models (ERGMs). In contrast to LSGMs, traditional formulations of ERGMs specify a model for a network through a joint distribution, by identifying particular global topological features to be included as statistics in the log-linear term of the joint distribution. Effects frequently included are edge density, transitivity, block effects, or covariate effects. Sets of dependent edge random variables and conditional distributions are induced by the terms included in the joint model, rather than explicitly defined as in a LSGM. Both the traditional formulations of ERGMs and LSGMs have joint distributions in Gibbsian form (Casleton et al., 2014b), but these joint distributions must be constructed for a LSGM from the set of specified conditional distributions (Kaiser and Cressie, 2000) which may be accomplished under certain conditions.

A LSGM is also an application of a binary Markov Random Field (MRF) model to the graph edges. This model, originally referred to as the auto-logistic model, was developed in Besag (1974) and is commonly used to analyze geo-referenced data due to its ability to model spatial dependence. Consider the conditional binary distribution written in exponential family form as

$$\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i)) = \exp[y(\mathbf{s}_i)A_i\{\mathbf{y}(N_i)\} - B_i\{\mathbf{y}(N_i)\}] \quad y(\mathbf{s}_i) = 0, 1$$

where $A_i\{\mathbf{y}(N_i)\}$ is referred to as the natural parameter function and $B_i\{\mathbf{y}(N_i)\}$ is a function of A_i, which for an auto-logistic model is $B_i\{\mathbf{y}(N_i)\} = \log[1 + \exp(A_i\{\mathbf{y}(N_i)\})]$. The natural

parameter function of Besag (1974) is, for $i = 1, \ldots, m$

$$A_i\{\mathbf{y}(N_i)\} = \alpha_i + \sum_{\mathbf{s}_j \in N_i} \eta_{ij} y(\mathbf{s}_j) \tag{5.1}$$

where α_i are leading constants and the η_{ij} are dependence parameters between pairs of random variables. This formula assumes pairwise-only dependence so that dependence is only modeled between pairs of dependent edges.

Let $S \equiv \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m\}$ be the collection of all edge markers and V as any subset of S. In order to incorporate higher-order dependence terms by explicitly modeling dependence between sets of random variables of size greater than two, Lee et al. (2001) presented a necessary form for the natural parameter function of the binary conditional distributions, for $i = 1, \ldots, m$, as

$$A_i\{\mathbf{y}(N_i)\} = \alpha_i + \sum_{V : \mathbf{s}_i \in V} \left[\theta_V \prod_{\mathbf{s}_j \in V \setminus \{\mathbf{s}_i\}} y(\mathbf{s}_j) \right] \tag{5.2}$$

where α_i are similar leading constants and θ_V dependence parameters between sets of random variables, which must be invariant to any permutation of the indices in V. Although this allows for dependence to be modeled between any sized set of dependent edges, all possible subsets of S are hardly ever considered. One reason results from the Hammersly-Clifford Theorem (Cressie, 1993, p. 417) which implies that $\theta_V = 0$ unless the edges in V represent a clique. A clique is a single random variable or any set of random variables such that all random variables in the set are neighbors of every other random variable in the set. Thus, the neighborhood definitions defined for the edges of the network play a large role in the terms of the natural parameter function which are included.

The parameterization of the natural parameter functions in (5.1) and (5.2) is often referred to as the original, or uncentered parameterization. This form has been shown to lead to interpretation issues with the parameters, particularly in separating the large and small scale model structures and when attribute information is included in the model. To allow a more uniform interpretation of parameters across reasonable levels of statistical dependence, Caragea and Kaiser (2009) proposed a centered parameterization of the natural parameter function for binary conditional distributions. For simplicity we will assume common dependence parameters for each size of clique and will adjust for potentially varying sizes of dependence structures

(Casleton et al., 2014b). The parameterization presented in Caragea and Kaiser (2009), which also assumes a pairwise-only dependence, can be written as

$$A_i\{\mathbf{y}(N_i)\} = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i}\left[\frac{\eta_2}{|N_i| + |N_j|}\right](y(\mathbf{s}_j) - \kappa_j) \tag{5.3}$$

where κ_i represents the large scale structures, η_2 the dependence between pairs of edges and $|N_\ell|$ the size of the neighborhood for edge $y(\mathbf{s}_\ell)$. The sum of neighborhood sizes allows for the θ_V parameters in (5.2) to be invariant to the permutation of the indices. The centered parameterization was extended by Casleton et al. (2014a) to include cliques of size three with resulting natural parameter function

$$\begin{aligned} A_i\{\mathbf{y}(N_i)\} = &\log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i}\left[\frac{\eta_2}{|N_i| + |N_j|}\right](y(\mathbf{s}_j) - \kappa_j) \\ &+ \sum_{\mathbf{s}_j,\mathbf{s}_k \in \mathcal{C}_i^3}\sum\left[\frac{\eta_3}{|\mathcal{C}_i^3| + |\mathcal{C}_j^3| + |\mathcal{C}_k^3|}\right](y(\mathbf{s}_j)y(\mathbf{s}_k) - \kappa_j\kappa_k) \end{aligned} \tag{5.4}$$

where $\mathcal{C}_i^3 = \{\mathbf{s}_j : \mathbf{s}_j, \mathbf{s}_i, \text{ and } \mathbf{s}_k \text{ are neighbors for some } k\}$ are all cliques of size three corresponding to the random variable $Y(\mathbf{s}_i)$ of size $|\mathcal{C}_i^3|$. Note that with a little algebra, both (5.3) and (5.4) do correspond to the necessary form of $A_i\{\mathbf{y}(N_i)\}$ in (5.2).

By definition, a clique of size three is a set of three edges for which each pair are neighbors. Potential cliques of size three are determined by the specification of neighborhoods. A common neighborhood definition in network analysis is incidence, where two potential edges are considered neighbors if they share a common node. Configuration of edges, or subgraphs (Frank and Strauss, 1986), which lead to cliques of size three for an incidence definition of dependence are 3-stars and triangles (see Figure 5.2). The dependence term in (5.4) may be partitioned based on the type of subgraph or only a particular subgraph can be modeled. For example, if only cliques of size three which correspond to triangles are considered, the natural parameter function can be written as

$$\begin{aligned} A_i\{\mathbf{y}(N_i)\} = &\log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i}\left[\frac{\eta_2}{|N_i| + |N_j|}\right](y(\mathbf{s}_j) - \kappa_j) \\ &+ \sum_{\mathbf{s}_j,\mathbf{s}_k \in \mathcal{T}_i}\left[\frac{\eta_3}{|\mathcal{T}_i| + |\mathcal{T}_j| + |\mathcal{T}_k|}\right](y(\mathbf{s}_j)y(\mathbf{s}_k) - \kappa_j\kappa_k) \end{aligned} \tag{5.5}$$

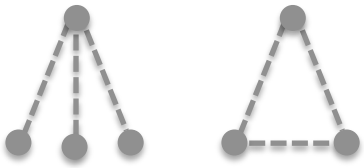


Figure 5.2 Subgraphs which correspond to cliques of size 3 given an incidence definition of dependence: a 3-star (left) and triangle (right).

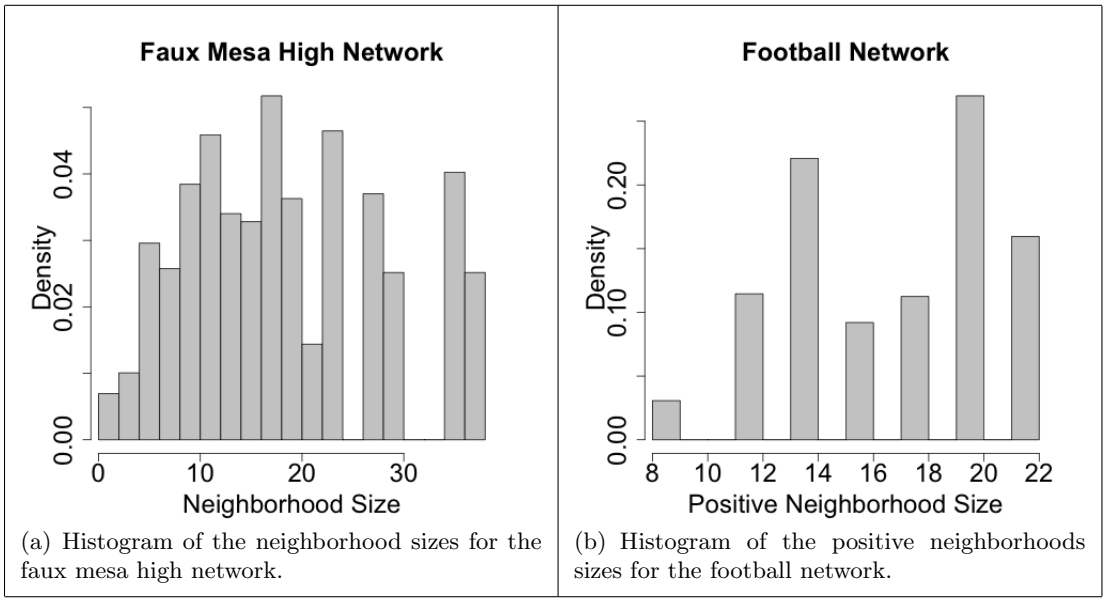
where $|T_\ell|$ is the number of potential triangles to which the edge $y(\mathbf{s}_\ell)$ belongs.

Lastly, it is not necessary to include all lower order dependence terms. Thus, a valid model will result from excluding the pairwise dependence term from the natural parameter functions in (5.4) and (5.5).

5.4 Exploratory Data Analysis of Networks

In this section, possible and realized topological features and dependence structures are contrasted between the Faux Mesa High and football networks. Recall there are 205 nodes in the Faux Mesa High network with 4,174 potential edges, and 115 nodes in the football network with 6,555 possible edges. Edges in the Faux Mesa High network are classified according to the potential gender pairing, so Female-Female, Male-Female or Male-Male edges. Conference play whether in- or out-of-conferences categorizes the edges of the football network.

The neighborhood definition for the Faux Mesa High network models dependence between two edges that share a node (incident) and connect the same gender and race pairing. For example, the neighborhood of an edge which potentially connects student A to student B, where both students are female and Hispanic, consists of all edges which connect student A to other Hispanic, females and student B to other Hispanic, females. Although this neighborhood definition may seem contrived, nodal covariates have been shown to be more important than network structures in predicting social network edge formation (Hunter et al., 2008a). The resulting neighborhood sizes are displayed in Figure 5.3 and have an average of 18.95.



(a) Histogram of the neighborhood sizes for the faux mesa high network.

(b) Histogram of the positive neighborhoods sizes for the football network.

Figure 5.3 Neighborhood sizes resulting from the neighborhood definitions of the Faux Mesa High and football networks.

Only edges which connect two schools in the same conference will be assigned a dependence structure in the football network. Games classified as out-of-conference will not have neighbors, and thus, will be modeled to form independently of other possible edges. This is intuitive because out-of-conference games typically are scheduled between individual schools and are not influenced by conference play. Of the 6,555 possible edges, 6,066 will not have neighbors. Two in-conference edges will be considered neighbors if they are incident and potentially join three schools within the same conference. Figure 5.3 displays neighborhood sizes for those edges which have at least one neighbor. The average positive neighborhood size is 17.12.

Edges which can possible form cliques of size three are a result of how the neighborhoods are defined. For the football network, the neighborhood definition leads to 12,507 possible cliques of size three, 1,395 of which are triangles. On average, an in-conference edge will be included in 76.73 cliques of size three and 8.56 triangles. Even with the restrictive neighborhood definition for the Faux Mesa High network, the number of potential cliques of size three is 162,229 of which an average edge is a member of 116.60. The number of cliques of size three which are triangles is only 3,093 (less than 2%) partly because triangles cannot form between edges that

Table 5.1 Structural comparison of the Faux Mesa High and football networks

	Football	Faux Mesa High
# nodes	115	205
# potential edges	6555	4174
Number of neighbor-less edges	6066	21
Average neighborhood size	17.12*	18.95
Average number of cliques of size 3	76.73*	116.60
Average number of triangles	8.56*	11.13*
Number of cliques of size 3	12,507	162,229
Number of triangles	1395	3093
Number of unique 2-stars	4185	39,546

*Average of only the positive values.

potentially connect dissimilar race (e.g., Hispanic–Black) or sex (e.g. Male-Female) pairings. A possible edge between the same race and sex can form 11.13 potential triangles, on average. A summary of the structures for both networks is displayed in Table 5.1.

Another way to compare the two networks is through the two-stars, or pairs of dependent edges in an incidence definition of dependence. The football network contains 4,185 unique, potential 2-stars, where the Faux Mesa High network contains 39,546. Note that every possible clique of size three contains three 2-stars (see Figure 5.2), but the same 2-star can compose many cliques of size three. Figure 5.4 summarizes how many cliques of size three each unique 2-star can potential form. For the football network, most 2-stars are contained in 10 cliques of size three, where the most common number for the Faux Mesa High network is 19. Note that the minimum number of cliques of size three to which a 2-star is contained is four for the football network, but there are a multiple 2-stars in the Faux Mesa High network which are not contained in any cliques of size three.

Restricting attention to only triangles, note that a potential 2-star cannot be part of more than one potential triangle (Figure 5.2). Although every possible triangle contains three 2-stars, a 2-star can be in at most one triangle. Table 5.2 summarizes the number of potential 2-stars which are in either zero or one triangle for both networks. A majority of 2-stars in the

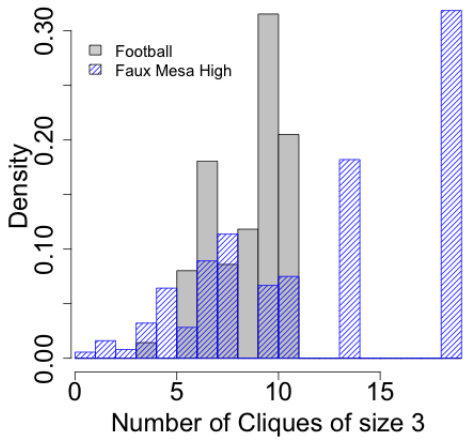


Figure 5.4 Histograms of the number of cliques of size three to which the unique 2-stars belong for both the Faux Mesa High and football networks.

Table 5.2 Number of potential 2-stars which can form either zero or one triangle.

	Number of triangles	
	0	1
Football	0	4185
Faux mesa high	30,267	9279

Faux Mesa High network cannot potentially form a triangle, where every 2-star in the football network can.

The previous discussion focused on the number of potential structures that each network contains, but not how many of those structures are realized in the actual network. Table 5.3 displays a summary of the realized structures in each network. About 10% of the 6,555 possible edges in the football network are realized, while only 3.9% of the 4,174 Faux Mesa High edges are realized. The proportion of edges realized is similar for all categories of the Faux Mesa High network, but there is a large discrepancy for the proportion of realized edges for out-of-conference games (0.03) and in-conference games (0.826) in the football network. A striking difference between the two networks is the percentage of topological features realized. In the Faux Mesa High network, less than 1% of the possible 2-stars, cliques of size three, and triangles are realized. In contrast, about half of all these structures are realized in the football network,

Table 5.3 Comparison of the realized structures of the Faux Mesa High and football networks.

	Football	Faux Mesa High
Overall Proportion	0.094	0.039
Category Proportions:	Out: 0.03	Male-Male: 0.035
	In: 0.826	Male-Female 0.027
		Female-Female: 0.067
Category Counts:	Out: 6066	Male-Male: 1047
	In: 489	Female-Male: 2085
		Female-Female: 1042
Realized 2-stars	2679 (64%)	126 (0.3%)
Realized cliques of size 3	5823 (47%)	68 (0.04%)
Realized triangles	731 (52%)	6 (0.2%)

a consequence of the large percentage of in-conference edges realized and the restriction that only this category of edges have dependence structures.

5.5 Analysis

In this section, three comparable models will be fit to and examined for both the Faux Mesa High and football networks. All models for a given network will have the same large-scale structure. For the Faux Mesa High network, three large-scale parameters are included to represent the different rate of edge formation between nodes of a specified gender, κ_{FF} for Female–Female (FF) edges, κ_{MM} for Male–Male (MM) edges, or of the opposite gender, κ_{FM} for Female–Male (FM) edges. One κ_{in} for edges that form between two schools from the same conference and one κ_{out} for games between two schools from different conferences will be used in the analysis of the football network.

The difference between the three models will be in how the small-scale structure is represented. Model 1 will assume pairwise-only dependence with the corresponding natural parameter function shown in (5.3). Both dependence between pairs and triples of dependent edges will be included in Model 3. For the football network, the natural parameter function corresponds to (5.4). Due to the importance placed on transitivity in social network analysis, only cliques of size three that are triangles will be modeled for the Faux Mesa High network. Thus, the

Table 5.4 Parameter estimates and 90% percentile parametric bootstrap confidence intervals for three models fit to the Faux Mesa High network.

Parameter	Model 1	Model 2	Model 3
κ_{FF}	0.063 (0.049, 0.078)	0.063 (0.048, 0.079)	0.060 (0.043, 0.078)
κ_{FM}	0.025 (0.020, 0.031)	0.027 (0.021, 0.033)	0.026 (0.020, 0.032)
κ_{MM}	0.033 (0.024, 0.043)	0.033 (0.023, 0.044)	0.032 (0.020, 0.044)
η_2	8.40 (4.155, 10.732)	–	7.15 (0.093, 10.273)
η_3	–	36.51 (29.10, 51.16)	24.19 (18.076, 42.028)

natural parameter function for the Model 3 fit to the Faux Mesa High network is shown in (5.5). Lastly, Model 2 for each network will correspond to Model 3, but without the pairwise dependence term.

Parameter estimates were obtained by maximizing the log-pseudolikelihood, which for a LSGM is

$$\log \mathrm{PL} = \sum_i \{y(\mathbf{s}_i) \log[p_i(N_i)] + (1 - y(\mathbf{s}_i)) \log[1 - p_i(N_i)]\}.$$

Confidence intervals are obtained through a percentile parametric bootstrap. For each model, 1000 simulations were obtained through a Gibbs sampler using a burn-in and thinning of 10,000 networks. Point estimates and confidence intervals for the Faux Mesa High are displayed in Table 5.4. All 1,000 simulations are represented in the confidence intervals of Model 1; however, the estimation algorithm failed to produce estimates for 14 simulations from Model 2 and 33 from Model 3 and thus, the given confidence intervals are based on 986 and 967 simulations, respectively. The estimates for the the κ parameters are relatively close to the realized proportions (see Table 5.3) with confidence intervals which are not too wide. Confidence intervals for dependence parameters do not contain zero and thus indicate a significant amount of dependence between pairs of neighboring edges and triples of edges which form triangles.

The estimates of the parameters for the three models fit to the football network are displayed in Table 5.5. Confidence intervals for Models 1 and 2 are based on 1,000 simulations, where the algorithm failed to converge for 161 simulations of Model 3, and thus, its confidence intervals are based on 839 simulated networks. Again, the estimates for the large-scale parameters

Table 5.5 Parameter estimates and 90% percentile parametric bootstrap confidence intervals for three models fit to the football network.

Parameter	Model 1	Model 2	Model 3
κ_{out}	0.034 (0.031, 0.038)	0.034 (0.031, 0.038)	0.034 (0.030, 0.038)
κ_{in}	0.830 (0.796, 0.863)	0.831 (0.794, 0.873)	0.832 (0.801, 0.861)
η_2	3.54 (-3.299, 7.354)	–	-142.23 (-209.16, 2.23)
η_3	–	3.97 (-1.75, 7.79)	127.79 (-1.75, 186.45)

seem reasonable given the realized proportions in the network (see Table 5.3), and confidence intervals appear to be symmetric and narrow around the estimates. Confidence intervals for the dependence parameters do not indicate strongly significant dependence and are not as symmetric, particularly those for Model 3.

One consideration when fitting network models is the issue of model degeneracy. This phenomena occurs when a model places all or most of its probability on only a few possible networks, and often on those that do not resemble the network of interest. This model failure has been widely studied in the network analysis literature for ERGMs (Handcock, 2003a), and has also been recognized in a more general class of models for interactive systems (Strauss, 1986) and associated with long-range dependence in Ising models (Snijders, 2002). A method suggested by Kaiser et al. (2012) to identify model degeneracy is to simulate from the fitted model and verify the simulated proportions are reasonable. Figure 5.5 displays the simulated overall and category-wise proportions from the simulations obtained from the three fitted models to both example networks as Normal quantile-quantile plots. For reference, the dashed horizontal line represents the proportion in the realized corresponding network.

The first row of Figure 5.5 summarizes simulations of the Faux Mesa High network obtained from the three fitted models. Simulations from Models 2 and 3 overestimate the overall and same gender proportions. This feature is not necessarily indicative of model degeneracy, as model degeneracy occurs when the simulations result in only a few possible networks. Rather, the overestimation indicates the models are not adequately describing a feature of the network. All three models are able to recreate the distribution of Female–Male proportions with the realized value near the center of all distributions. The neighborhood definition does not enable

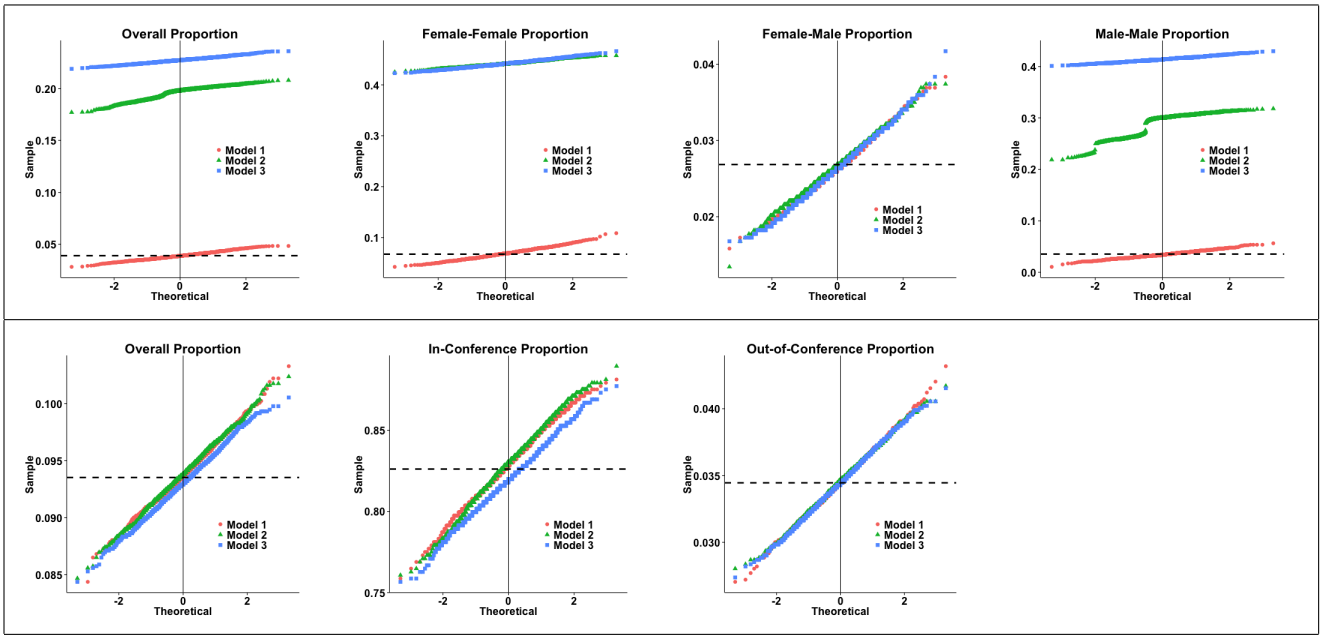


Figure 5.5 Normal quantile-quantile plots of the different proportions from the simulations from the three models fits to both network. The first row represents simulations from the fit to the Faux Mesa High network and the second row to the football network. The dashed horizontal line represents the proportion from the realized network. A vertical line at the theoretical quantile of zero has been drawn for reference.

potential triangles between these type of edges indicating the feature not adequately described by Models 2 and 3 is transitivity. Further, Model 1 which does not include a term for transitivity is more adequately able to recreate all proportions.

The second row of Figure 5.5 displays the simulated proportions from the three fitted models to the football network. Out-of-conference edges form independently under all three models, and thus the simulated proportions are nearly identical regardless of the model fit. Proportions are also similar between the three models for the in-conference, and thus overall, proportions with the middle of the distributions aligning with the realized proportion. Thus, none of the models raise concerns of model degeneracy or inadequacy, even though the estimates from Model 3 do not seem intuitive.

As a method of model interpretation and to further explore their behavior, the conditional expectations resulting from the fitted models will be explored. Conditional and marginal expectations are displayed for fitted Models 1 and 3 to the Faux Mesa High network in Table 5.6. Results of Models 2 and 3 are similar, thus expectations from Model 2 are not included. Because conditional expectations depend on the value and number of neighbors, expectations will be computed for the for the focal edge, represented by as a dashed line, in the possible configurations displayed in Figure 5.6. Although the particular configurations are not necessarily common in the Faux Mesa High network, they are used because they are simple to visualize. Marginal expectations are approximately the estimate of the corresponding κ because of the centered parameterization and only depend on the sex of the two nodes it would join. The potential configuration on the left of Figure 5.6 is used to compute the conditional expectations an edge connecting nodes of the same sex. This edge has four neighbors and can potentially form two triangles. Because no triangles can form between potential Female–Male edges, a comparable configuration is displayed on the right of Figure 5.6 where the focal edge has four neighbors.

For both Models 1 and 3 and for each type of edge, as the number of neighbors realized increases so does the conditional expectation. When zero, one or two neighbors which do not form a triangle are realized, the conditional expectations for same-sex edges are similar between the two models. For Female–Male edges, the two models result in similar conditional expectations for all possible neighborhood formations. However, if the realization of the focal edge will form a triangle, it will most certainly be realized under Model 1. When the large η_3 term is not included, for either Model 1 or Female–Male edges, the largest conditional expectation is only 0.858.

To compare the fit of the three models to the football network, the conditional expectations of a typical in-conference edge are computed. Conditional expectations depend on the number of positive neighbors, but for Models 2 and 3, also on the resulting number of positive cliques of size three. Twenty is the most common number of neighbors, and for each of these edges neighbors also have twenty neighbors. Thus, the denominator of the pairwise dependence term will be 40. These edges belong to 100 cliques of size three, as do the other 2 edges in

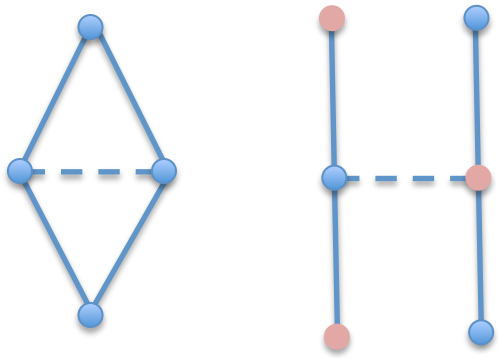


Figure 5.6 Possible configurations used to compute the conditional expectations for Female–Female and Male–Male (left) and Female–Male (right) in Table 5.6. The focal edge for which the conditional expectation is computed is the dashed line in both.

	Female–Female	Male–Male		Female–Male
Model 1				
Marginal Expectation, $E[y(\mathbf{s}_i)]$	0.063	0.033		0.025
Conditional Expectation, $E[y(\mathbf{s}_i)\|\mathbf{x}(\mathbf{s}_i), \mathbf{y}(N_i)]$				
0 neighbors realized	0.047	0.029		0.023
1 neighbor realized	0.142	0.089		0.071
2 neighbors realized	0.354	0.245		0.203
3 neighbors realized	0.645	0.519		0.458
4 neighbors realized	0.858	0.781		0.737
Model 3				
Marginal Expectation, $E[y(\mathbf{s}_i)]$	0.060	0.032		0.026
Conditional Expectation, $E[y(\mathbf{s}_i)\|\mathbf{x}(\mathbf{s}_i), \mathbf{y}(N_i)]$				
0 neighbors realized	0.046	0.028		0.023
1 neighbor realized	0.117	0.074		0.072
2 neighbors realized				0.203
From same potential triangle	0.994	0.989		
From different potential triangle	0.269	0.180		
3 neighbors realized	0.998	0.996		0.456
4 neighbors realized	1	1		0.734

Table 5.6 Marginal and conditional expectations for edges with the potential configuration shown in Figure 5.6.

the clique, leading to a denominator of the three-way dependence term of 300. The resulting conditional expectations for all three models are plotted in Figure 5.7 against the number of positive neighbors, $\{0, 1, \ldots, 20\}$, along with the approximate marginal expectation from each model as a dashed, horizontal line.

The conditional expectations for Model 1 increase monotonically with the number of positive neighbors. Because Models 2 and 3 also include the dependence from the triples of edges, there are multiple possibilities for a given number of positive neighbors. For example, if three neighbors are positive, the number of cliques of size three which are positive could be either 1, 2, or 3. Conditional expectations for Models 1 and 2 are similar, particularly once the number of positive neighbors is 7, possibly resulting from the similarity of the estimates of η_2 and η_3 in each model. This also demonstrates that the effect of few neighbors realized is greater than the effect of few cliques of size three realized since the minimum conditional expectation for Model 1 is 0.53 and 0.66 for Model 2.

Conditional expectations from Model 3 in Figure 5.7 is not intuitive as this value is almost 1 when few neighbors are realized and the expectation decreases once half of the neighbors are realized. This is due to the fact that the estimate of η_2 is negative and it is not until about 75% of the neighbors are realized that the dependence from the cliques of size three begin to overwhelm the effect from the pairwise dependence on the conditional expectation.

Lastly, the three models are examined as to how well they are able to recreate certain structures of the networks. Figure 5.8 demonstrates the three models ability to recreate the number of realized 2-stars and corresponding dependent triples of edges that were modeled. The top two plots show the results from the models fit to the Faux Mesa High network. Simulations from the fitted Models 2 and 3 result in too many 2-stars and triangles, which is intuitive as the edges in general from these models were over simulated (see Figure 5.5). Model 1 does not include a term that explicitly models transitivity; however, it most able to recreate the number of realized triangles. The number of triangles from simulations of Model 1 do tend to be lower than expected as 401 of the 1,000 simulations result in none of the potential triangles realized. However, six of the simulations produce the correct number of triangles where the closest number of triangles in a simulation from Model 3 was off by 3,058 triangles. Thus, the

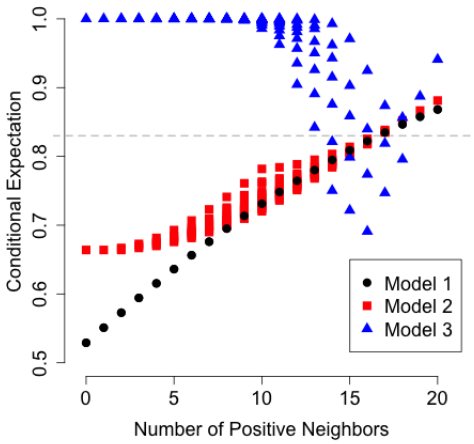


Figure 5.7 Conditional expectations for the three models based on number of positive neighbors. Approximate marginal expectation for each model is plotted as a gray, dashed horizontal line.

Faux Mesa High network does not contain enough realized transitivity to be able to estimate its effect with a parameter. Although transitivity has a natural interpretation for social networks, friends of friends are more likely friends, the data does not support the inclusion of a parameter which accounts for it.

The use of a saturated graph and the particular neighborhood definition in the Faux Mesa High network are what results in only 3,093 possible triangles, 6 or 0.2% of which are realized. Consider the lack of a saturated graph where all $\binom{205}{2} = 20,910$ potential edges are modeled. Further, allow neighborhoods induced by common ERGMs, so that each edge is neighbors with all $2(205-2) = 406$ incident edges. This would result in 1,414,910 possible triangles, of which 62 (0.004%) are realized. So, although it could be argued that the saturated graph and neighborhoods are arbitrarily defined, their use is not what contributes to the lack of model fit, but rather the lack of realized transitivity in the Faux Mesa High network.

The ability of the three models to reproduce the 2-stars and cliques of size three in the football network is displayed in the second row of Figure 5.8. Again, all three models behave similarly. The simulations tend to over recreate both the 2-stars and cliques of size three of

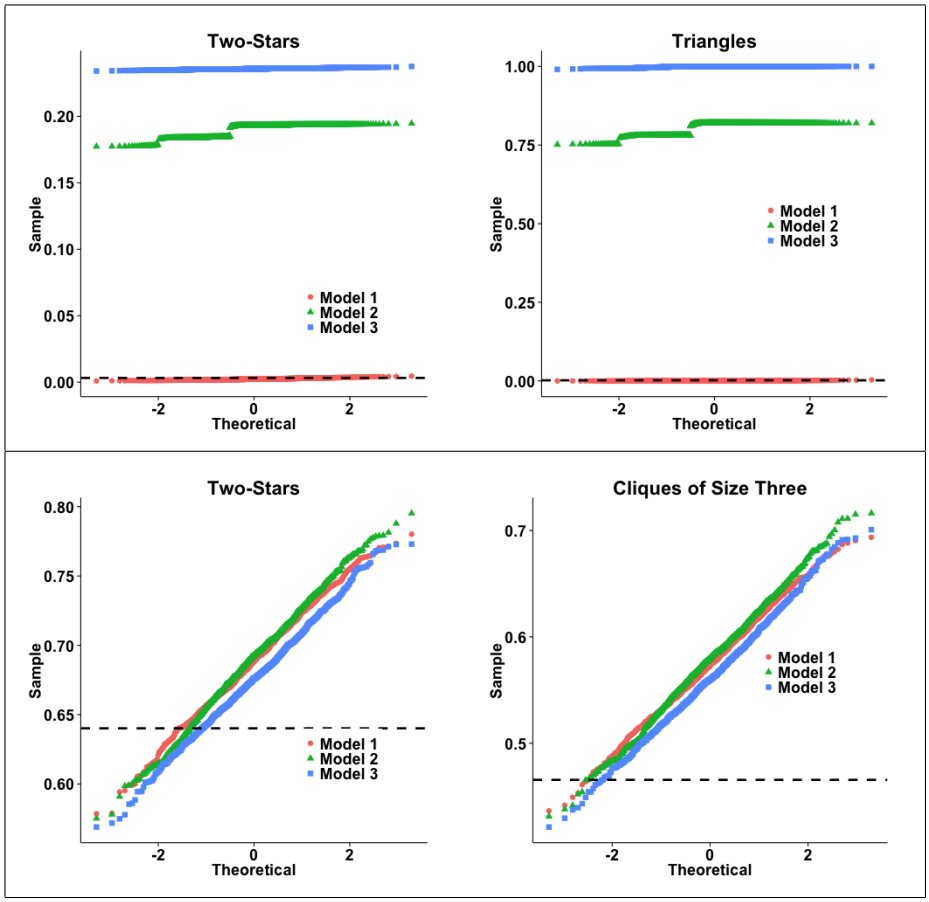


Figure 5.8 Normal quantile-quantile plots which demonstrate the ability of the three models to recreate the 2-stars and triples of dependent edges modeled in each of the Faux Mesa High and football networks. Vertical, dashed lines correspond to the actual

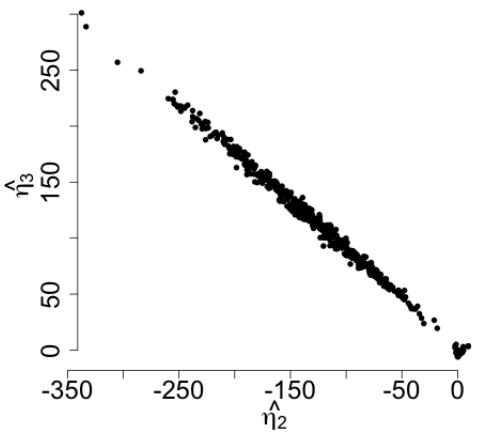


Figure 5.9 Scatterplot of the estimates of η_2 against η_3 for the 839 simulations of Model 3 to the football network.

the network, on average. Thus, it may be that the strength of the dependence between pairs and triples or dependent edges is slightly overestimated in all three models.

Table 5.2 of Section 5.4 shows that all 2-stars of the football network are members of at least four cliques of size three, and that every 2-star can potentially form a triangle. Including both a term which captures the dependence between pairs of edges and a term for triples of edges may not be necessary. This is further demonstrated in Figure 5.9 which plots the estimate of η_2 against the estimate for η_3 in each of the 839 simulations of Model 3. Both parameters are attempting to describe the same feature of the data, as the correlation coefficient between the two sets of estimates is -0.997. As to which model (Model 1 or Model 2) is more appropriate for this particular network is a modeling choice that depends on the type of dependence, pairwise or triples, is more important to describe.

5.6 Discussion

The previous sections describe and fit LSGMs to two structurally different example networks. The first, the Faux Mesa High network, is sparse with few topological features of interest realized, making it difficult to estimate parameters which describe these features. Although

transitivity is an important feature of social networks, it may be the case that the network of interest does not provide enough information to adequately assess the effect, as was seen with the Faux Mesa High network. The football network has a large percentage of multiple structures realized, but due to an overlap of these features, a model which attempts to describe both features is unable to separate the effects.

These two example networks also provide a cautionary tale to blindly fitting network models. If Model 3 had been fit to the Faux Mesa High network and only the estimates, bootstrap estimates, and resulting confidence intervals had been examined, the model would potentially not have been questioned. It is when the simulations from this model are examined that it becomes clear that the model is not adequately describing a feature of interest. Similarly, an adequate model fit could have been determined if only Model 3 had been fit to the football network and only the simulations had been examined. Conditional expectations indicate a nonintuitive model fit which was further verified by the colinearity in the η_2 and η_3 estimates from the simulations. These two examples demonstrate that it is important to have a good understanding of the network before attempting to interpret or assess the model fit.

Although the difficulty in fitting network models is demonstrated with the LSGM, a similar argument can be made for other network analysis approaches. In fact, a LSGM is an alternative method of specifying an ERGM (Casleton et al., 2014b), and thus the results apply to the more widely-used joint specification of ERGMs.

CHAPTER 6. GENERAL CONCLUSIONS

6.1 General Discussion

Network analysis is hard.

Networks have become a popular modeling tool because many applications can be conceptualized as a network. They can represent complex dependencies, are equipped to handle massive and high-dimensional data, can incorporate attribute information, and provide an attractive approach to visualizing data. However, development of the ability to draw meaningful conclusions from a network has lagged behind.

The existing literature on network analysis is quickly growing and spread across many different disciplines. Vivar and Banks (2012) cite a difficulty in the analysis of a network is that most networks require a "specific, ad hoc model", and that it is unlikely there will ever be a "unified theory" for network models. Another complication is how to match questions of interest with an appropriate manner in which to answer them. For example, centrality, a notion of the importance of individual nodes, has been identified as an important topic, but there is no agreement on which of the large number of proposed measures can appropriately capture this feature (Kolaczyk, 2010). Chakrabarti and Faloutsos (2006) details the seven most popular methods to calculate the exponent on the power law of a degree distribution, one of the three important features graph generation algorithms are to replicate.

This dissertation introduces a new class of models for network analysis, local structure graph models (LSGMs). Development was motivated by the binary Markov Random Field (MRF) models, which were first introduced as the auto-logistic model of Besag (1974). Even though MRF models have been frequently used in the analysis of spatial data, their application to

networks is not immediate. To appropriately apply the MRF models to networks requires the inclusion of higher-order dependence, as well as an extension of the centered parameterization.

The LSGM approach does not claim to solve a unifying theory for all network analysis, nor is it appropriate for every realized network. The advantages of this modeling approach is an interpretable and controllable local dependence structure and an ability to separate the large and small scale model structures. The interpretability advantage is particularly important when attribute information is available about the network. Thus, a LSGM is an appropriate model for a network with local structure that can be used to describe the global behavior of the network. However, the use of a LSGM, or any network analysis model should always be accompanied by careful model diagnostics.

6.2 Recommendation for Future Research

There are two important additions to the class of LSGMs related to the common issue of model degeneracy.

The main requirement in specifying a LSGM is the definition of neighborhoods, which defines the dependence structure of the model. Although specifying this structure provides more control over it, as opposed to ERGMs for which the structure is induced, this also implies that a modeling decision must be made as to what sets of edges will be modeled as dependent. Connected to the neighborhood definition is the use of a saturated graph, an optional modeling feature that restricts the number of edges to be modeled. This restriction affects the neighborhoods as there are less neighborhoods to define, and it can help decrease the size of neighborhoods.

In the examples of this dissertation, the choice of neighborhoods and saturated graph were application-specific decisions, often made by taking into account features thought to effect the dependence of the edges. The choice of neighborhoods is related to model degeneracy because the model breaks down if the neighborhoods become too large, also identified by Schweinberger and Handcock (2012) for ERGMs. What defines a "large" neighborhood may also be application specific. More generally, model degeneracy occurs when the small-scale structure of the model overwhelms the large-scale structure. The large-scale should represent

overall, global features of the network and the small-scale local deviations from the overall pattern. In practice this corresponds to global homogeneity with local heterogeneity. If a fitted model is degenerate, it can sometimes be attributed to too much structure in the small-scale component of the model and an adjustment to adding more of the structure to the large-scale component will avoid degeneracy. Thus, the first addition to LSGMs is to formulate a more structured approach to specifying the saturated graph and neighborhoods in a manner which aligns with the application, but also takes into consideration model degeneracy.

The second addition to the LSGM modeling approach is a metric to identify when the model has become degenerate. Currently, the recommendation, from both the MRF model and ERGM literature, is to simulate from the fitted model and verify that the simulations are able to replicate features of the observed network. From MRF models, the recommendation is to check the simulated proportion of realized edges and ERGM suggest checking the topological features included in the model specification. However, checking the simulations does not always indicate a problem, as demonstrated by the analysis of the football network in Chapter 5. In addition, the boundary between the areas of the parameter space where the model is and is not degenerate is not well defined. Therefore, there exists an area where the model is not fully degenerate, but is not appropriately describing the data of interest. The development of a metric to assess the degree of model degeneracy would be a beneficial further technique for the LSGM.

BIBLIOGRAPHY

Agee, E., Snow, J., and Clare, P. (1976). Multiple vortex features in the tornado cyclone and the occurrence of tornado families. *Monthly Weather Review*, 104(5):552–563.

Aiello, W., Chung, F., and Lu, L. (2001). A Random Graph Model for Power Law Graphs. *Experimental Mathematics*, 10(1):53–66.

Arnold, B. C. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156.

Bar, S., Gonen, M., and Wool, A. (2007). A geographic directed preferential internet topology model. *Computer Networks*, 51(14):4174–4188.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barigozzi, M., Fagiolo, G., and Garlaschelli, D. (2010). The multi-network of international trade: A commodity-specific analysis. Working paper, Laboratory of Economics and Management, Sant'Anna School of Advanced Studies.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. John Wiley & Sons.

Ben-Avraham, D., F Rozenfeld, A., Cohen, R., and Havlin, S. (2003). Geographical embedding of scale-free networks. *Physica A: Statistical Mechanics and its Applications*, 330(1):107–116.

Bender-deMoll, S., Morris, M., and Moody, J. (2008). Prototype packages for managing and animating longitudinal network data: `dynamicnetwork` and `rSoNIA`. *Journal of Statistical Software*, 24(7).

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36(2):192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24:179–195.

Besag, J. (1992). Contribution to the discussion of Geyer, C.J. and E.A. Thompson, constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):657–699.

Besag, J. (2001). Markov chain Monte Carlo for statistical inference. *Center for Statistics and the Social Sciences*.

Bhamidi, S., Bresler, G., and Sly, A. (2008). Mixing time of exponential random graphs. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 803–812. IEEE Computer Society.

Buell, M. F., Buell, H. F., and Small, J. A. (1971). Invasion of trees in secondary succession on the New Jersey Piedmont. *Bulletin of the Torrey Botanical Club*, pages 67–74.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55.

Caragea, P. C. and Kaiser, M. S. (2009). Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(3):281–300.

Casleton, E. M., Kaiser, M. S., and Nordman, D. J. (2014a). Local structure graph models with higher-order dependence. In Preparation.

Casleton, E. M., Nordman, D. J., and Kaiser, M. S. (2014b). A local structure model for network analysis. *Statistics and Its Interface*. Submitted.

Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys*, 38(1).

Clifford, P. (1990). Markov random fields in statistics. *Disorder in physical systems*, pages 19–32.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley-Interscience, New York.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Fienberg, S. E. (2012). Brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839.

Frank, O. and Strauss, D. (1986). Markov Graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):657–699.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Gill, P. S. and Swartz, T. B. (2004). Bayesian analysis of directed graphs data with applications to social networks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(2):249–260.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233.

Goodreau, S. M. (2007). Advances in Exponential Random Graph (p*) Models Applied to a Large Social Network. *Social networks*, 29(2):231–248.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., and Morris, M. (2008). A statnet tutorial. *Journal of statistical software*, 24(9):1.

Goodreau, S. M., Kitts, J. A., and Morris, M. (2009). Birds of a Feather, Or Friend of a Friend?: Using Exponential Random Graph Models to Investigate Adolescent Social Networks. *Demography*, 46(1):103–125.

Groendyke, C., Welch, D., and Hunter, D. (2012). A network-based analysis of the 1861 hagelloch measles data. *Biometrics*.

Gross, J. L. and Yellen, J. (2006). *Graph theory and its applications*. CRC press.

Gumpertz, M. L., Graham, J. M., and Ristaino, J. B. (1997). Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(2):pp. 131–156.

Guo, J., Wilson, A. G., and Nordman, D. J. (2013). Bayesian nonparametric models for community detection. *Technometrics*, 55(4):390–402.

Guyon, X. (1995). *Random fields on a network: modeling, statistics, and applications*. Springer.

Handcock, M. S. (2003a). Assessing degeneracy in statistical models of social networks. Working Paper 39, Center for Statistics and the Social Sciences, University of Washington, Seattle.

Handcock, M. S. (2003b). Statistical models for social networks: Inference and degeneracy. *Dynamic social network modeling and analysis*, 126:229–252.

Handcock, M. S. and Gile, K. J. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Bender-deMoll, S., and Morris, M. (2014). *statnet: Software tools for the Statistical Analysis of Network Data*. The Statnet Project (http://www.statnet.org). R package version 2014.2.0.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1–11.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society*, 170(2):301–354.

Hoff, P. D. (2003). Random effects models for network data. In *Dynamic social network modeling and analysis: Workshop summary and papers*, pages 303–312. National Academies Press, Washington, DC.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Hoff, P. D. and Ward, M. D. (2005). Analyzing dependencies in international relations: commerce, capitalism, conflict, cooperation, and democracy. In *46th Annual Convention of the International Studies Association*, pages 1–20, Honolulu, HI.

Holland, P. and Leinhardt, S. (1981). An Exponential family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Hughes, J., Haran, M., and Caragea, P. C. (2011). Autologistic models for binary data on a lattice. *Environmetrics*, 22(7):857–871.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social networks*, 29(2):216–230.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008a). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258.

Hunter, D. R. and Handcock, M. S. (2006). Inference in Curved Exponential Family Models for Networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008b). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29.

Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4):856–882.

INSNA (2013). International network for social network analysis.

Kaiser, M. S., Caragea, P. C., and Furukawa, K. (2012). Centered parameterizations and dependence limitations in Markov random field models. *Journal of Statistical Planning and Inference*, 142(7):1855–1863.

Kaiser, M. S. and Cressie, N. (2000). The construction of multivariate distributions from Markov random fields. *Journal of Multivariate Analysis*, 73(2):199–220.

Kashirin, V. and Dijkstra, L. (2013). A heuristic optimization method for mitigating the impact of a virus attack. *Procedia Computer Science*, 18:2619–2628.

Kolaczyk, E. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer.

Kolaczyk, E. (2010). Tutorial: Statistical analysis of network data. In *2010–11 Program on Complex Networks Opening Tutorials & Workshop*. SAMSI.

Koskinen, J. H., Robins, G. L., and Pattison, P. E. (2010). Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology*, 7(3):366–384.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks*, 31(3):204–213.

Kuhn, F., Moscibroda, T., and Wattenhofer, R. (2004). Unit disk graph approximation. In *DIALM-POMC*, pages 17–23, Philadelphia, Pennsylvania.

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):1–5.

Lee, J., Kaiser, M. S., and Cressie, N. (2001). Multiway dependence in exponential family conditional distributions. *Journal of Multivariate Analysis*, 79(2):171–190.

Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. (2010). Kronecker Graphs : An Approach to Modeling Networks. *Journal of Machine Learning Research*, 11:985–1042.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2.

Lofgren, E. (2012). Visualizing results from infection transmission models. *Epidemiology*, 23(5):738–741.

Lubbers, M. J. and Snijders, T. (2007). A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. *Social Networks*, 29(4):489–507.

Lunga, D. and Kirshner, S. (2011). Generating similar graphs from spherical features. In *Ninth Workshop on Mining and Learning with Graphs (MLG '11)*, San Diego, CA.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(Suppl 2):S186–S188.

Morris, M., Handcock, M. S., and Hunter, D. R. (2008). Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. *Journal of Statistical Software*, 24(4):1548–7660.

Neumayer, S. and Modiano, E. (2010). Network reliability with geographically correlated failures. In *Proceedings IEEE INFOCOM*, pages 1–9. IEEE.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(1):2566–72.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Onat, F. and Stojmenovic, I. (2007). Generating random graphs for wireless actuator networks. In *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, pages 1–12.

Park, J. and Newman, M. E. (2004). Solution of the two-star model of a network. *Physical Review E*, 70(6):066146.

Park, J. and Newman, M. E. (2005). Solution for the properties of a clustered network. *Physical Review E*, 72(2):026136.

Pattison, P. and Robins, G. (2002). Neighborhood–based models for social networks. *Sociological Methodology*, 32(1):301–337.

Pattison, P. and Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–193.

Pickett, S. (1982). Population patterns through twenty years of oldfield succession. *Vegetatio*, 49(1):45–59.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., et al. (1997). Protecting adolescents from harm: findings from the national longitudinal study on adolescent health. *Jama*, 278(10):823–832.

Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, 3:446–484.

Robins, G., Pattison, P., and Elliott, P. (2001). Network models for social influence processes. *Psychometrika*, 66(2):161–189.

Robins, G., Pattison, P., and Wasserman, S. (1999). Logit models and logistic regressions for social networks: Iii. valued relations. *Psychometrika*, 64(3):371–394.

Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29(2):192–215.

Salter-Townshend, M., White, A., Gollini, I., and Murphy, T. B. (2012). Review of statistical network analysis: models, algorithms, and software. *Statistical Analysis and Data Mining*, 5(4):243–264.

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370.

Schweinberger, M. and Handcock, M. S. (2012). Hierarchical exponential-family random graph models with local dependence.

Schweinberger, M., Petrescu-Prahova, M., and Vu, D. Q. (2012). Disaster response on September 11, 2001 through the lens of statistical network analysis. Technical report, Department of Statistics, Pennsylvania State University.

Simpson, S. L., Hayasaka, S., and Laurienti, P. J. (2011). Exponential random graph modeling for complex brain networks. *PloS ONE*, 6(5):e20039.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Snijders, T. A. B. (2007). Contribution to the discussion of Handcock, M. S., A. E. Raftery, and J. M. Tantrum, Model-based clustering for social networks. *Journal of the Royal Statistical Society*, 170(2):301–354.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36:99–153.

Sporns, O., Chialvo, D. R., Kaiser, M., Hilgetag, C. C., et al. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425.

Strauss, D. (1986). On a General Class of Models for Interaction. *SIAM Review*, 28(4):513–527.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212.

Van Duijn, M. A., Snijders, T. A., and Zijlstra, B. J. (2004). p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2):234–254.

Vivar, J. C. and Banks, D. (2012). Models for networks: a cross-disciplinary science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(1):13–27.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic Blockmodels for Directed Graphs. *Journal of the American Statistical Association*, 82(397):8–19.

Wasserman, S. and Anderson, C. (1987). Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, 9(1):1–36.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks. *Psychometrika*, 61(3):401–425.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Wong, G. Y. (1987). Bayesian models for directed graphs. *Journal of the American Statistical Association*, 82(397):140–148.

Zhu, J., Huang, H.-C., and Wu, J. (2005). Modeling spatial-temporal binary data using Markov random fields. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2):212–225.

Zijlstra, B. J., Duijn, M. A., and Snijders, T. A. (2009). MCMC estimation for the p2 network regression model with crossed random effects. *British Journal of Mathematical and Statistical Psychology*, 62(1):143–166.